Geophysical Research Letters

Supporting Information for

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[1,3], 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

Contents of this file

Table S1 and Figures S1 to S2, with corresponding descriptions

Introduction

This supporting information provides a table, figures, and descriptions that are supplementary to the main article.

1. Reference Datasets

In the paper, we use the same observations as in Planton et al. (2021), and refer to these as our default reference datasets, as listed in Table S1. In addition, the following observation-based reference datasets were used as alternative datasets to measure the range of observational discrepancies (shown in Fig. 2 and S1): 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 and Arkin, 1997), ERA-20C (Poli et al., 2016), ECMWF Reanalysis v5 (ERA-5, Hersbach et al., 2020), Hadley Centre 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., 2020).

2. Inter-member Spread

Figure S1 is based on the same statistics used in Figure 1 of the manuscript, but without normalization. The circles in each panel represent the average error across all members as compared to our reference data (Table 2), with vertical line markers showing the results for individual members. These plots collectively illustrate the inter-model skill differences, the inter-member (internal) variability in the errors for each model. For some metrics, notably those based on mean state characteristics, the inter-member spread due to internal variability is very narrow. The *Double ITCZ Bias, Equatorial Precipitation, SST, and Taux Biases* (Figs. S1a-d), have particularly small spread among members. The internally-generated spread is much larger for *ENSO Amplitude* (Fig. S1e), *ENSO Duration* (Fig. S1f), *ENSO Asymmetry* (Fig. S1j), and Ocean-driven SST (Fig. S1x), where some members nearly match the observations while others differ strongly from observed (e.g., CanESM2 for *ENSO Asymmetry*; Fig. S1j); for these metrics, multiple members are needed to obtain an accurate assessment of skill relative to observations. Figure S1 also shows that the inter-member spread is model dependent. For *ENSO Duration* (Fig. 2f), the inter-member spread is much larger for CanESM5 than CESM1-CAM5, despite a similar number of ensemble members.

3. Monte-Carlo Sampling for Pseudo-ensembles

To estimate the ensemble size needed to gauge ENSO performance, we apply a Monte Carlo approach as proposed by Milinski et al. (2020). For each large ensemble 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). The means of PEs from the "without replacement" sampling results converges to a single scalar value, the full ensemble mean, with increasing N, since for $N = N_{full}$ each PE is identical to the original full ensemble. On the contrary, means of PEs from the "with replacement" sampling does not converge to the mean when the entire sample size is considered.

In Fig. S2, random samples were generated as described above. The dashed orange (blue) lines represent the 5% to 95% range of the pseudo-ensemble means from the "with (without) replacement" sampling. The without replacement sampling results converging to the full ensemble mean (black line). In Figure 3 of the main article, the average magnitude in the error distributions of each panel of Fig. S2 is shown, along with an estimate of a minimum ensemble size needed to resolve differences in skill between the models.

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, as shown in Fig. 3 of the main article. The criteria we selected is different to the study of Milinski et al. (2020), where they selected 5% from "without replacement" sampling. In their study, considering a substantially larger ensemble size (N=200) was used which was only available for one model, there was less concern that N_{min} to be biased low when the "without replacement" sampling was used. However in our study, using a suite of multi-models but individuals have 20-65 ensemble members, we use the "with replacement" sampling for defining N_{min} because it approximates what would happen if the samples had been drawn from the underlying infinite-member distribution. Additional comparison of these sampling approaches can be found in the Appendix of Milinski et al. (2020).

Disclaimer

Table S1. Reference datasets used in the CLIVAR 2020 ENSO metrics package. Monthly means, interpolated to a 1° × 1° grid, are used for each dataset.

Dataset	Field(s)	Epoch	Reference
AVISO	Sea surface height (SSH)	1993–2018	AVISO (see Data Availability Statement)
ERA-Interim	Surface temperature (TS)	1979–2018	Dee et al. (2011)
GPCPv2.3	Precipitation (PR)	1979–2018	Adler et al. (2003)
TropFlux	Sea surface temperature (SST), Net heat flux (NHF), Latent heat flux (LHF), Sensible heat flux (SHF), Longwave radiation (LWR), Shortwave radiation (SWR), Zonal wind stress (Taux)	1979–2018	Praveen Kumar et al. (2012, 2013)

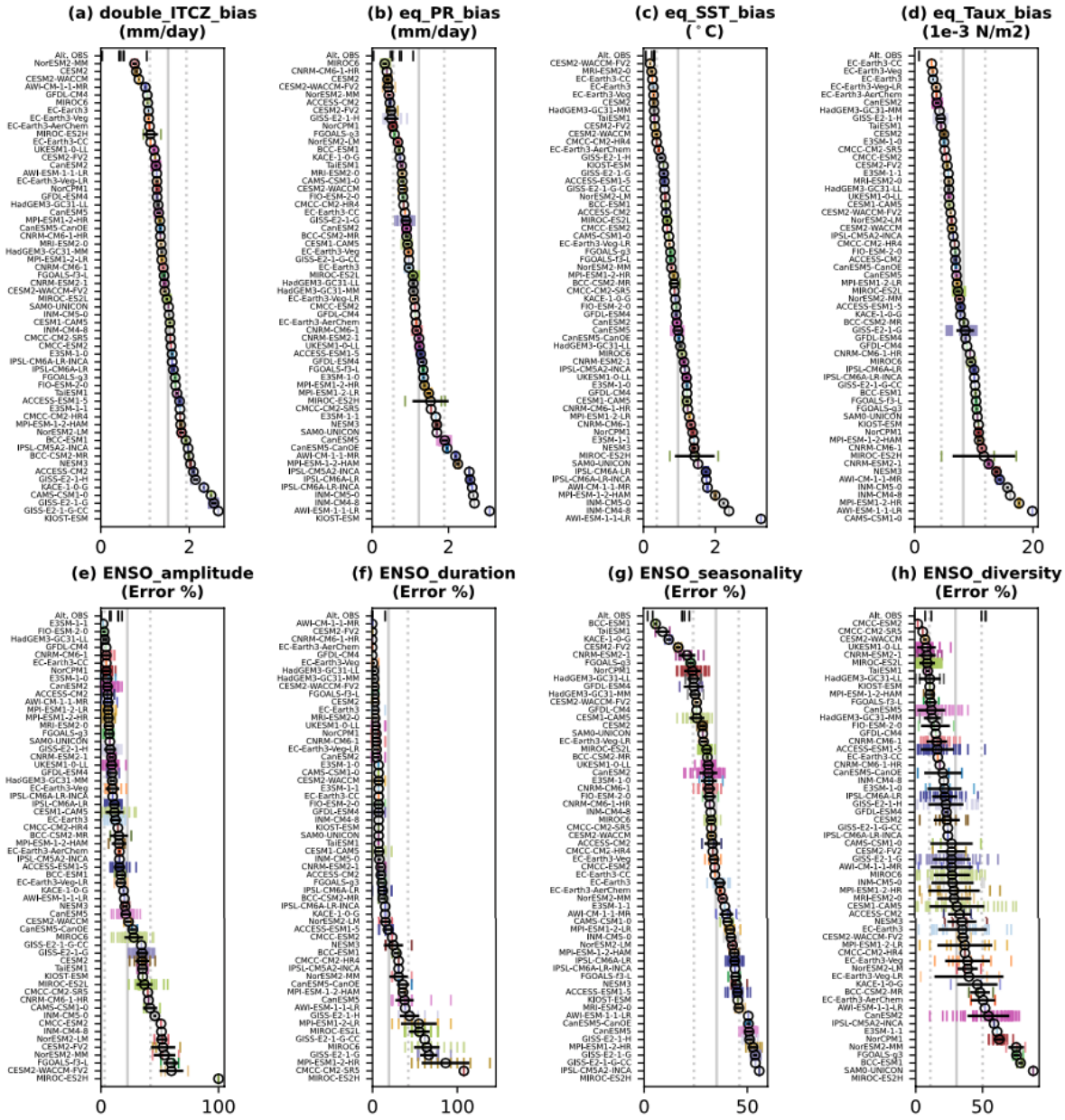

Figure S1. Error metrics calculated for ensemble simulations from the CMIP6 models and Large ensembles of CESM1-CAM5 and CanESM2. Lines represent standard deviation of error metrics for individual ensemble members, with circles denoting the average of ensemble members for any given model. Metrics Collection categories include *Performance* (panels a to o), *Teleconnections* (panels p to s), and *Processes* (panels t to x). 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. In some panels there are models without metrics values because some variables were not available or problems with the model output were identified.

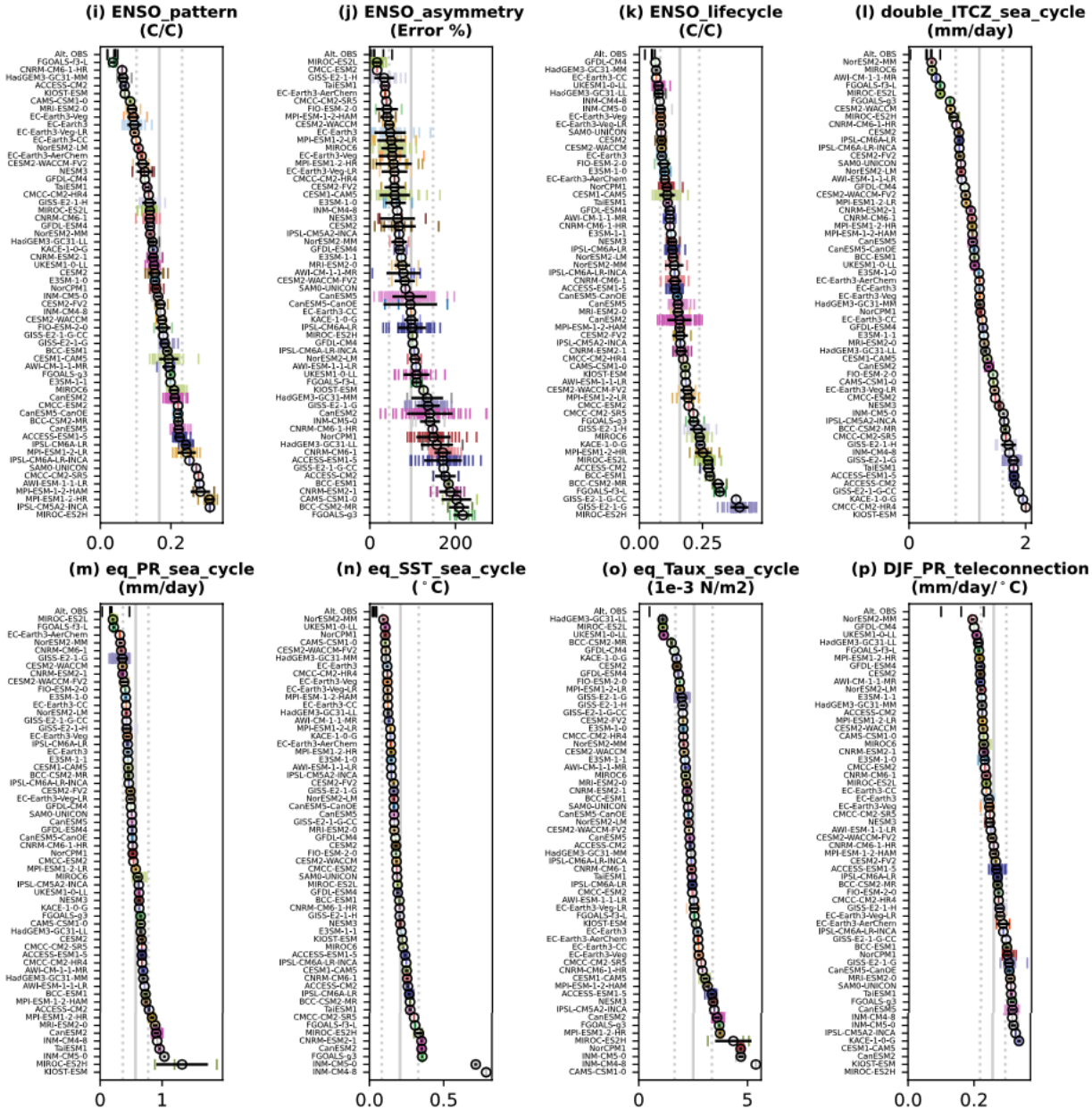

Figure S1. (continued, #1)

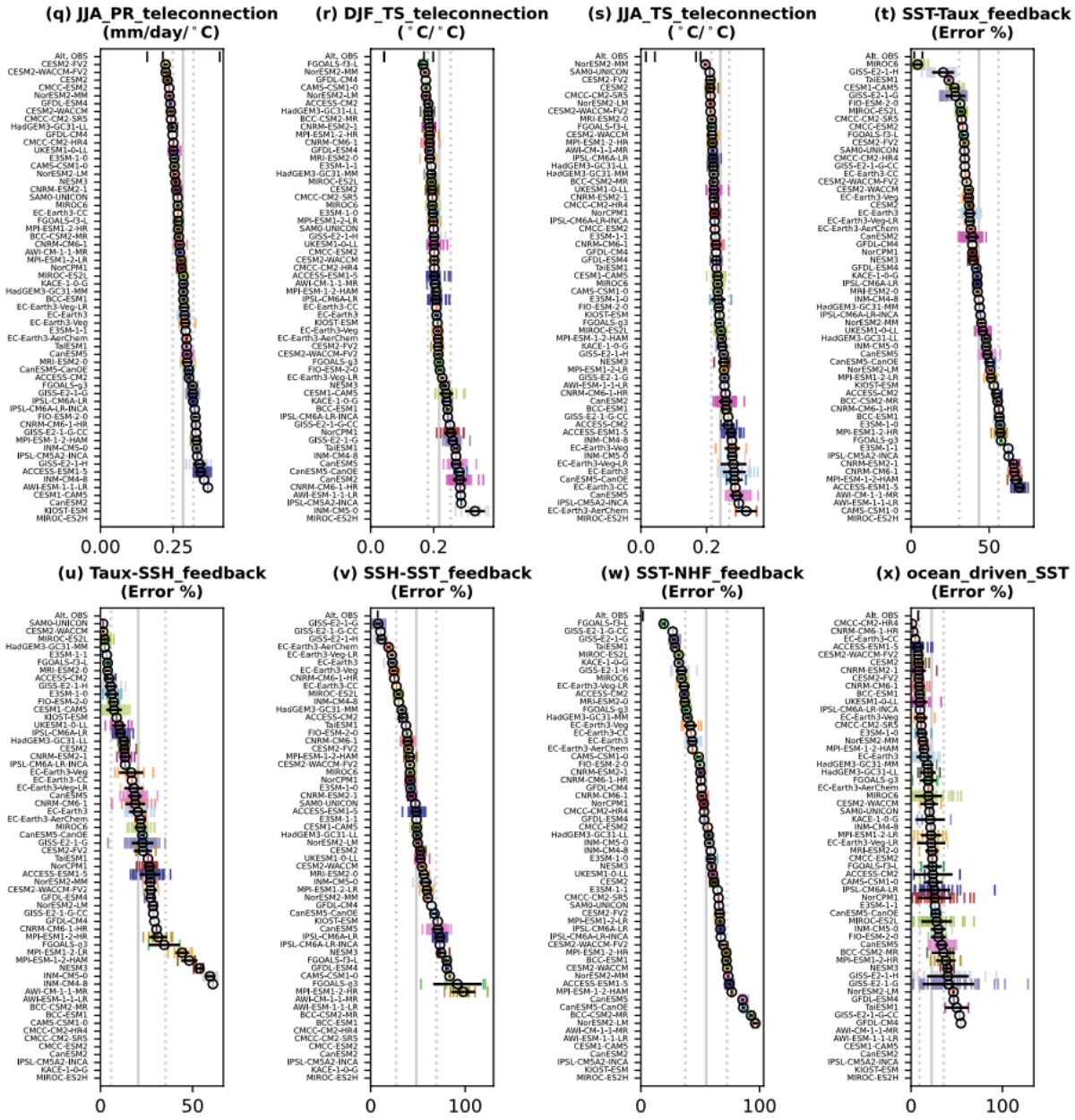

Figure S1. (continued, #2)

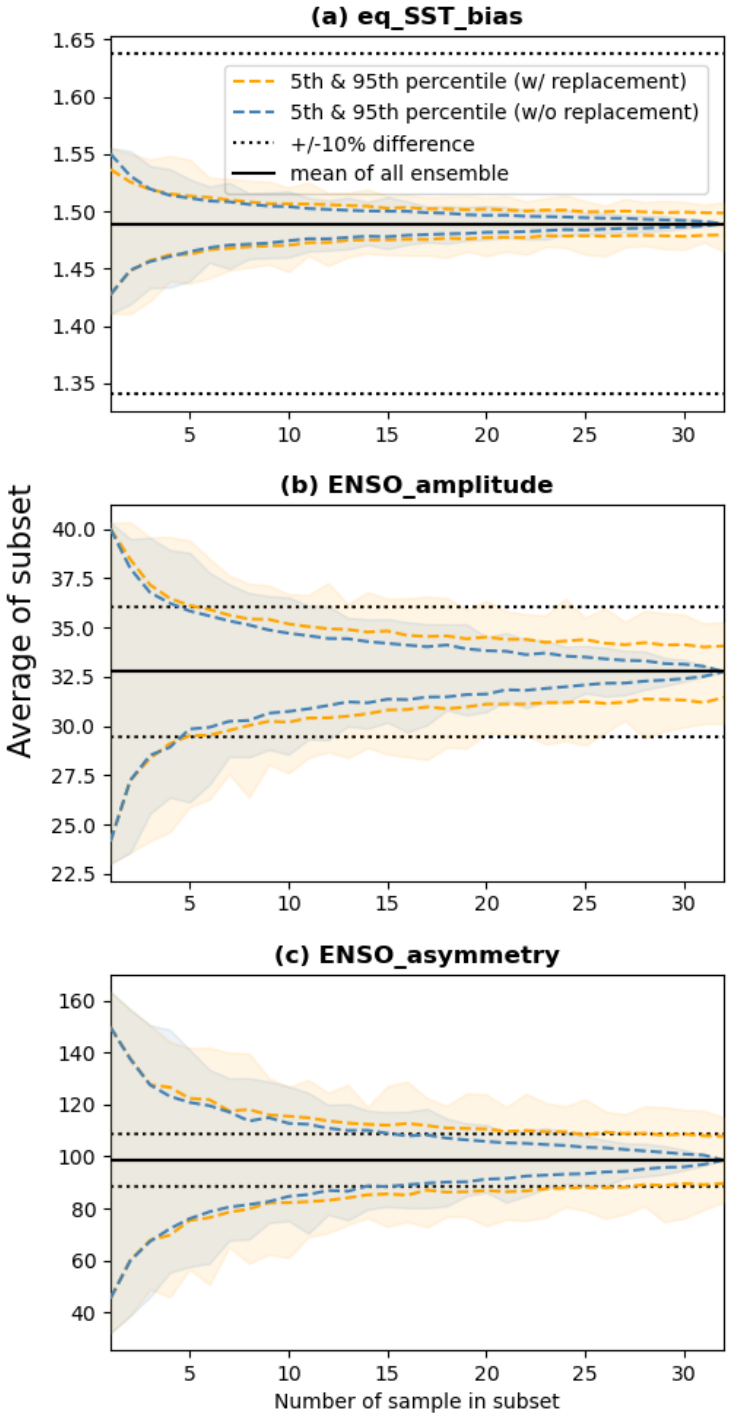

Figure S2. Distribution of the sample mean for pseudo-ensembles from IPSL-CM6A-LR ensemble with (orange) or without (blue) replacement samplings at different sample sizes (abscissa). Three representative metrics are shown: (a) *Equatorial SST Bias*, (b) *ENSO Amplitude* and (c) *Asymmetry*. Shaded area indicates the full min-max range of the sample distribution, long-dashed lines indicate 5th and 95th percentiles of the sample distribution, and short-dashed lines indicate a difference of 10% from the mean of the full ensemble.