JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2023MS004147

Estimating Uncertainty in Simulated ENSO Statistics

Yann Y. Planton[1,2], Jiwoo Lee[3], Andrew T. Wittenberg[4], Peter J. Gleckler[2], Éric Guilyardi[5,6], Shayne McGregor[1,7], and Michael J. McPhaden[2]

[1]School of Earth Atmosphere and Environment, Monash University, Clayton, VIC, Australia, [2]NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA, [3]Lawrence Livermore National Laboratory, Livermore, CA, USA, [4]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, NJ, USA, [5]LOCEAN-IPSL, CNRS-IRD-MNHN-Sorbonne Université, Paris, France, [6]NCAS-Climate, University of Reading, Reading, UK, [7]ARC Centre of Excellence for Climate Extremes, Monash University, Clayton, VIC, Australia

Abstract Large ensembles of model simulations are frequently used to reduce the impact of internal variability when evaluating climate models and assessing climate change induced trends. However, the optimal number of ensemble members required to distinguish model biases and climate change signals from internal variability varies across models and metrics. Here we analyze the mean, variance and skewness of precipitation and sea surface temperature in the eastern equatorial Pacific region often used to describe the El Niño–Southern Oscillation (ENSO), obtained from large ensembles of Coupled model intercomparison project phase 6 climate simulations. Leveraging established statistical theory, we develop and assess equations to estimate, a priori, the ensemble size or simulation length required to limit sampling-based uncertainties in ENSO statistics to within a desired tolerance. Our results confirm that the uncertainty of these statistics decreases with the square root of the time series length and/or ensemble size. Moreover, we demonstrate that uncertainties of these statistics are generally comparable when computed using either pre-industrial control or historical runs. This suggests that pre-industrial runs can sometimes be used to estimate the expected uncertainty of statistics computed from an existing historical member or ensemble, and the number of simulation years (run duration and/or ensemble size) required to adequately characterize the statistic. This advance allows us to use existing simulations (e.g., control runs that are performed during model development) to design ensembles that can sufficiently limit diagnostic uncertainties arising from simulated internal variability. These results may well be applicable to variables and regions beyond ENSO.

Plain Language Summary Earth's climate naturally fluctuates on intraseasonal to interdecadal timescales, confounding the evaluation of climate models and the detection of trends linked to climate change. To tackle this challenge, scientists produce ensembles of simulations with identical external forcings (e.g., volcanic eruptions, greenhouse gas emissions) but plausibly different initial conditions. In this study, we analyze how these ensembles can be used to reduce the uncertainty of the simulated climate, to help guide the design of future ensembles and optimize the use of available computing resources.

Key Points:

- Large ensembles of climate simulations are analyzed to characterize the sampling uncertainty of El Niño–Southern Oscillation statistics
- As expected, uncertainty of these statistics decreases with the square root of the number of simulated years provided by the ensemble
- A simple equation developed to predict the ensemble size required to limit this sampling uncertainty to within a given tolerance

Supporting Information:

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

Correspondence to:

Y. Y. Planton,
yann.planton@monash.edu

Citation:

Planton, Y. Y., Lee, J., Wittenberg, A. T., Gleckler, P. J., Guilyardi, É., McGregor, S., & McPhaden, M. J. (2024). Estimating uncertainty in simulated ENSO statistics. *Journal of Advances in Modeling Earth Systems*, *16*, e2023MS004147. https://doi.org/10.1029/2023MS004147

Received 29 NOV 2023
Accepted 28 AUG 2024

Author Contributions:

Conceptualization: Yann Y. Planton, Jiwoo Lee, Andrew T. Wittenberg, Shayne McGregor, Michael J. McPhaden
Formal analysis: Yann Y. Planton, Jiwoo Lee
Investigation: Yann Y. Planton
Methodology: Yann Y. Planton, Andrew T. Wittenberg, Shayne McGregor
Software: Yann Y. Planton

1. Introduction

The El Niño–Southern Oscillation (ENSO) is the largest source of interannual climate variability on the planet (see McPhaden et al., 2020 for a review), affecting the global atmospheric circulation (Taschetto et al., 2020), severe weather (Goddard & Gershunov, 2020), wildfire activity (Chen et al., 2017), agriculture (Anderson et al., 2018), fisheries (Bertrand et al., 2020), and economic activity (Cashin et al., 2017). ENSO is characterized by a recurring climate pattern involving a warming (El Niño) or a cooling (La Niña) of the sea surface temperature (SST) in the central and eastern tropical Pacific Ocean. The pattern shifts back and forth irregularly every two to seven years, with SST anomalies (SSTA) typically between 1°C and 3°C (Kestin et al., 1998).

Climate models are a primary tool for improving our understanding of Earth's past, present and future climate. Knowing how well climate models represent key aspects of the historical climate, and in particular ENSO variability, is critical for both further model development and to build trust in the model's ability to simulate past and future climate. Multiple phases of the Coupled Model Intercomparison Project (CMIP; Eyring et al., 2016; Meehl et al., 2000, 2007; Taylor et al., 2012) have enabled the benchmarking of climate models performance across development cycles, as well as identifying the relative strengths and weaknesses of each model. ENSO has been particularly scrutinized from one phase of the project to another (AchutaRao & Sperber, 2006; Bellenger

Supervision: Andrew T. Wittenberg, Peter J. Gleckler, Éric Guilyardi, Shayne McGregor, Michael J. McPhaden
Visualization: Yann Y. Planton
Writing – original draft: Yann Y. Planton
Writing – review & editing: Yann Y. Planton, Jiwoo Lee, Peter J. Gleckler, Éric Guilyardi, Shayne McGregor, Michael J. McPhaden

et al., 2014; Planton et al., 2021), highlighting, for example, a reduction of mean state biases and an improvement of the representation of ENSO variability.

Earth's climate naturally fluctuates on intraseasonal to interdecadal timescales (hereafter "internal variability"), which reduces our ability to detect projected ENSO changes with global warming (e.g., Maher et al., 2018; Ng et al., 2021; Wittenberg, 2009; Zheng et al., 2018) as well as robustly evaluating model performance (J. Lee et al., 2021). The use of model ensembles (each ensemble is created by starting simulations using a given model configuration from different initial conditions) is an established approach to identify the impact of internal variability on model characteristics and projections (e.g., Deser et al., 2020).

Due to the computational expense of running ensembles, modeling centers contributing to CMIP typically produce a limited number of ensemble simulations (i.e., fewer than 10 members). However, several studies indicate that 30 to 50 members may be required to robustly characterize ensemble mean decadal-scale trends of SST variance (Maher et al., 2018; Milinski et al., 2020; J. Lee et al., 2021). These 3 papers reached their conclusions by analyzing several very large ensembles and randomly selecting members of an ensemble to indicate how many members are required to obtain a given confidence interval on the ensemble mean. This random selection-based method is sophisticated but limited by the existing ensemble. In addition, it is somewhat complicated for those who simply need to estimate the Required Ensemble Size (RES) for a given expected uncertainty.

The epoch length used to perform an analysis is of utmost importance. Indeed, Cai et al. (2022) demonstrate that the lack of consensus about whether ENSO amplitude will increase with climate change in the Intergovernmental Panel on Climate Change Sixth Assessment Report (IPCC AR6; J.-Y. Lee et al., 2021) can be explained by the short 20-year epoch used. By using 100-year epochs, Cai et al. (2022) show that ~80% of the models (only one member per model is used) indicate an increase of ENSO amplitude, depending on the scenario. Doing so, they argue that with longer epochs the uncertainty of the statistic decreases. But a systematic approach hasn't been undertaken yet.

Thompson et al. (2015) proposed that a pre-industrial control run (piControl) provides a robust estimate of the simulated internal variability and therefore a single member per model is needed. This approach assumes that the internal variability is not changing with climate change, and that this single member is close to the center of the distribution (as the confidence interval is centered on the ensemble mean). Nevertheless, if the internal variability in piControl and historical runs are similar, one could use the piControl run to estimate a priori the number of members to compute for the historical run.

In this study, we employ established statistical theory to propose a complementary approach for estimating the RES for an expected uncertainty. We compare theoretical and empirical results relative to statistical uncertainty and test how reliable it is to use a piControl run to estimate the simulated internal variability. It follows that, provided that a long simulation is available, our results deliver new information about the ensemble uncertainty before the ensemble is generated, enabling those who perform the experiments to estimate a priori the number simulations to be performed, given a level of accuracy needed for a particular application. We provide equations to compute the uncertainty of the ensemble mean of a given ensemble (Equation 9) or to estimate the ensemble size required to reach a given uncertainty of the ensemble mean (Equation 10), without having to compute random selections. This yields a framework to quantify how the uncertainty of the ensemble mean is affected by the ensemble size (Section 3.1) and by the epoch length used to compute a statistic (Section 3.2). After comparing the uncertainty of the ensemble mean in piControl and historical runs (Section 3.3), we provide test cases using our equations making it possible for others to estimate the ensemble size for their own applications (Section 3.4).

2. Data and Methods

2.1. Model Simulations and Observations

We use piControl and historical runs from the model intercomparison project phase 6 (CMIP6) (Eyring et al., 2016). The historical runs, which aim to simulate the observed climate, are forced by time-varying natural (e.g., orbital parameters, solar irradiance and volcanic aerosols) and anthropogenic (e.g., aerosols and greenhouse gas emissions, and land use) forcings that are based on observations (e.g., Durack et al., 2018). In the piControl run, which is designed to simulate the unforced variability arising from processes internal to the climate system, natural and anthropogenic forcings are fixed to their estimated 1,850 values. We use 59 ensembles from 53

Table 1
List of Coupled Model Intercomparison Project Phase 6 Ensembles, Their Duration for piControl Run and Size for Historical Run

Model name	Ensemble	PI	HI	Model name	Ensemble	PI	HI
ACCESS-CM2	i1p1f1	500	10	GFDL-ESM4	i1p1f1	500	3
ACCESS-ESM1-5	i1p1f1	1,000	40	**GISS-E2-1-G_p1f1**	i1p1f1	851	12
AWI-CM-1-1-MR	i1p1f1	500	5	**GISS-E2-1-G_p1f2**	i1p1f2	851	11
BCC-CSM2-MR	i1p1f1	600	3	GISS-E2-1-G_p3f1	i1p3f1	601	9
BCC-ESM1	i1p1f1	451	3	GISS-E2-1-G_p5f1	i1p5f1	501	9
CAMS-CSM1-0	i1p1f1	500	2	**GISS-E2-1-H_p1f1**	i1p1f1	801	10
CanESM5_p1	i1p1f1	1,000	25	GISS-E2-1-H_p1f2	i1p1f2	451	5
CanESM5_p2	i1p2f1	1,051	40	GISS-E2-1-H_p3f1	i1p3f1	451	5
CanESM5-1	i1p1f1	501	47	GISS-E2-2-G	i1p3f1	351	5
CanESM5-CanOE	i1p2f1	501	3	**HadGEM3-GC31-LL**	i1p1f3	1,350	55
CESM2	i1p1f1	1,201	11	HadGEM3-GC31-MM	i1p1f3	500	4
CESM2-FV2	i1p1f1	500	3	INM-CM4-8	i1p1f1	531	1
CESM2-WACCM	i1p1f1	499	3	**INM-CM5-0**	i1p1f1	1,201	10
CESM2-WACCM-FV2	i1p1f1	500	3	**IPSL-CM6A-LR**	i1p1f1	2,000	33
CIESM	i1p1f1	500	3	MCM-UA-1-0	i1p1f1	500	1
CMCC-CM2-SR5	i1p2f1	500	10	MIROC-ES2H	i1p4f2	420	3
CMCC-ESM2	i1p1f1	500	1	**MIROC-ES2L**	i1p1f2	500	30
CNRM-CM6-1	i1p1f2	500	29	**MIROC6**	i1p1f1	800	50
CNRM-CM6-1-HR	i1p1f2	300	1	MPI-ESM-1-2-HAM	i1p1f1	1,000	3
CNRM-ESM2-1	i1p1f2	500	11	**MPI-ESM1-2-HR**	i1p1f1	500	10
E3SM-1-0	i1p1f1	500	5	**MPI-ESM1-2-LR**	i1p1f1	1,000	50
E3SM-2-0	i1p1f1	500	21	**MRI-ESM2-0**	i1p1f1	701	10
EC-Earth3	i1p1f1	1,105	18	NESM3	i1p1f1	500	5
EC-Earth3-AerChem	i1p1f1	501	3	**NorCPM1**	i1p1f1	1,500	30
EC-Earth3-CC	i1p1f1	505	10	NorESM2-LM	i1p1f1	501	3
EC-Earth3-Veg	i1p1f1	500	10	NorESM2-MM	i1p1f1	500	3
EC-Earth3-Veg-LR	i1p1f1	501	3	SAM0-UNICON	i1p1f1	700	1
FGOALS-f3-L	i1p1f1	561	3	TaiESM1	i1p1f1	500	2
FGOALS-g3	i1p1f1	700	6	**UKESM1-0-LL**	i1p1f2	1,880	16
GFDL-CM4	i1p1f1	500	1				

Note. Model ensembles considered as LEs are bolded. The member column indicates the fixed initialization procedures (i), physical parameterizations (p), and forcings (f) used for the ensemble. If several ensembles are available, the varying parameter is added to the model's name. The piControl column (PI) indicates the duration of the run, in years. The historical column (HI) indicates the number of members. Ensembles available as of October 2023. Further information on each model at https://es-doc.org/cmip6/.

models for which both historical and piControl runs are available and the piControl run is at least 300 years long (see Table 1 for the list of ensembles and their size). Monthly means are used for all data sets.

We consider 26 of these ensembles as "large ensembles" (LEs) as they give access to 10 samples or more for both piControl and historical runs (for more details about members and ensembles see Text S1 in Supporting Information S1). The 10 samples threshold enables a good balance between sample size (available number of historical members and duration of the piControl experiment; see Section 2.2.2 for details about the creation of distributions) and number of ensembles used in the study. A multi-model ensemble (MME; hereafter CMIP6-MME) is created using the first member of each 59 ensembles.

Note that we performed a simple quality control procedure: (a) we computed piControl's global mean surface temperature to verify if the simulated climate is stationary; and (b) we compared the diagnostics (defined in Section 2.2.1) computed from piControl and the corresponding historical runs to verify if the climate statistics are similar. Following this quality control, simulations of CAS-ESM2-0 and KACE-1-0-G are not used in this study, and the first 650 years of HadGM3-GC31-LL's piControl are also not used (for more details see Text S2 and Figure S1 in Supporting Information S1).

The epoch 1985–2014 of two observations data sets are used, Global Precipitation Climatology Project Monthly Analysis Product version 2.3 for precipitation (PR) (GPCPv2.3; Adler et al., 2003) and NOAA Optimum Interpolation Sea Surface Temperature version 2 for SST (OISSTv2; Reynolds et al., 2002).

2.2. Methodology

2.2.1. Diagnostics

It is common to compute the mean, variance, and skewness of a record to describe respectively our climate's mean state, variability and asymmetry (e.g., the fact that El Niño events can reach larger amplitudes than La Niña events). For a record of n time steps, the sample mean ($\overline{x}$), variance (σ^2) and skewness (g_1) can be defined as follows (e.g., Cramér, 1946):

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{2}$$

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2\right)^{2/3}} \tag{3}$$

Figure 1 illustrates the difficulty of evaluating and ranking models using the observed and modeled 30-year (1985–2014) mean, variance, and skewness of PR and SST (interannual anomalies are used for variance, and skewness) computed over the region Niño3 (hereafter N3; 90–150°W, 5°S–5°N), a key region for ENSO. The model ensemble from the CMIP Phase 6 (CMIP6) ensemble (59 different ensembles; red boxplots) displays a large range of values around the observations (horizontal black lines). If we compare the range of the CMIP6 MME to that of the single-model initial condition ensemble (made of 33 Historical simulations of IPSL-CM6A-LR model; purple boxplots), it is evident that internal variability has a considerable impact on PR skewness (Figure 1j), as well as SST variance and skewness: the IPSL-CM6A-LR ensemble covers 50% or more of the CMIP6 ensemble (Figures 1h and 1l). In this case, the range of values taken by the IPSL-CM6A-LR model would not strongly impact the evaluation of the model as the distance to the observation is large (e.g., Figures 1b, 1d, 1f, 1j, and 1l). However, if the ensembles are not very large, comparing two models in terms of N3 SSTA variance or skewness may not be conclusive (Figures 1h and 1l). Similarly, only relatively large changes in these statistics may be robustly detected in climate projections.

The mean ($\overline{x}$; Equation 1), variance (σ^2; Equation 2) and skewness (g_1; Equation 3) of N3 averaged PR and SST are analyzed. To do so, the domain average is computed, then the time series are analyzed using epoch lengths ranging from 30 to 150 years (every 15 years, i.e., 30, 45, 60, etc.). Each epoch is analyzed independently, the linear trend is removed (computed over the given epoch) and, for the variance and skewness, the seasonal cycle is removed (computed over the given epoch). These calculations are done using the CLIVAR ENSO metrics package (Planton et al., 2021), executed via the PCMDI Metrics Package framework (Lee et al., 2024).

The N3 region was selected to illustrate the results as it is often used in the literature (e.g., Jin et al., 2020; Yun et al., 2021) and has a positive skewness for both PRA and SSTA (Figures 1i and 1j). However, the results are generally true in regions Niño3.4 (120–170°W, 5°S–5°N) and Niño4 (160°E−150°W, 5°S–5°N) (not shown).

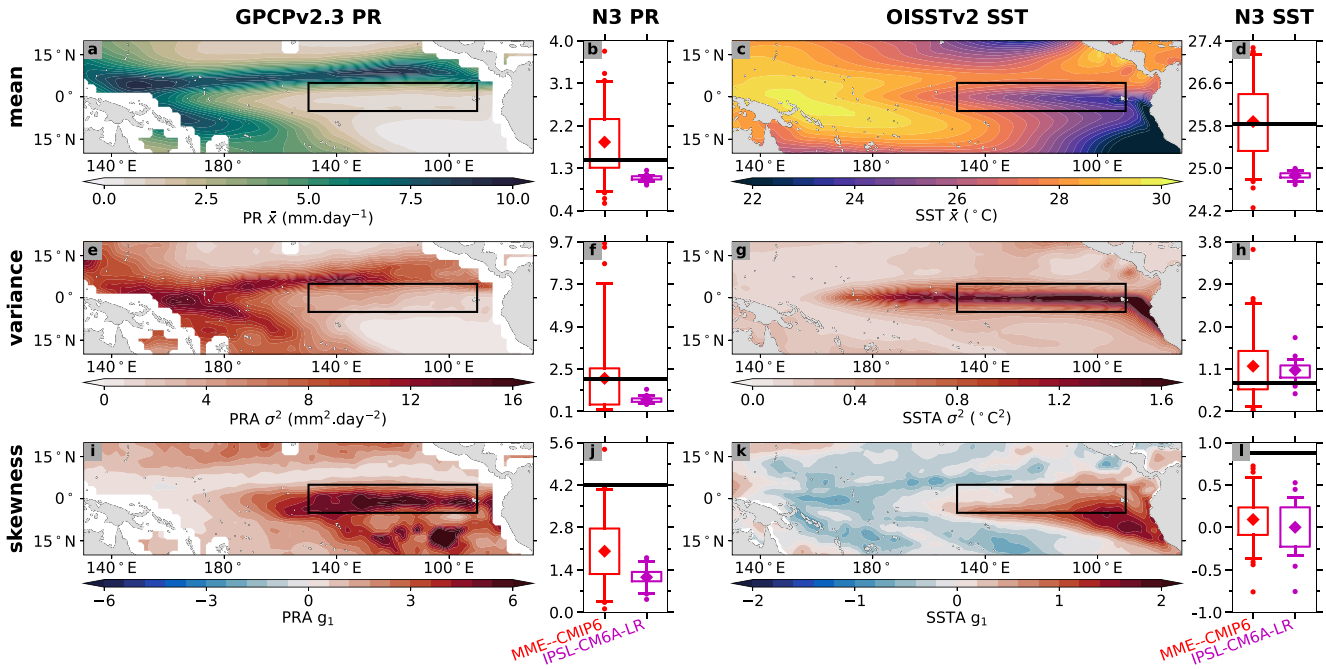

Figure 1. Statistical moments computed with observed and modeled precipitation (PR) and sea surface temperature (SST). Maps of observed PR (a, e, i; left column) and SST (c, g, k; right column) over the tropical Pacific Ocean, alongside Niño3 averaged (black rectangle) modeled (boxplots) and observed (black line) PR (b, f, j) and SST (d, h, l). Statistical moments are: mean (Equation 1; first row), variance (Equation 2; second row) and skewness (Equation 3; third row). The epoch 1985–2014 is used for all data sets. Boxplots represent the distributions of statistics computed from a multi-model ensemble (MME; 59 model intercomparison project phase 6 ensembles, red) and a single-model ensemble (33 IPSL-CM6A-LR members described in Boucher et al., 2020; purple). Whiskers extend to the 5th and 95th percentiles; boxes encompass the 25th and 75th percentiles; a diamond marks the mean; and dots indicate values that fall outside the whiskers.

Additional figures are available on the github webpage of the paper (https://yyplanton.github.io/estimating_uncertainties_enso/).

2.2.2. Creating piControl and Historical Distributions

For piControl runs, time series (Figure 2a) are split into non-overlapping epochs of a given epoch length (e.g., 30-year and 60-year; Figures 2b and 2d). The statistic is then computed on each epoch independently and all the statistical values are grouped into a single distribution per epoch length (Figures 2c and 2e). The overall length of the piControl run and the epoch length used to compute the statistics influence the number of values in piControl distributions: for a 300-year long run, 10 values will be available using 30-year epochs, but only 2 using 150-year epochs.

For historical ensembles, distributions are created using members with identical initialization procedure, physics and forcing (see Text S1 in Supporting Information S1 for more details). Time series (Figure 2f) of each member is split into epochs of a given epoch length (e.g., 1850–1879 is the first 30-year epoch). The statistic is then computed on each member independently and the statistical values from all members at the given epoch are grouped into a distribution (Figures 2g and 2h). This process is then repeated by moving forward by 5 years (e.g., 1855–1884 is the second 30-year epoch) until the end of the historical time series. All these distributions (e.g., there are 28 30-year epochs staggered by 5 years within the historical run 1850–2014) are used to describe an historical ensemble. The intra-ensemble mean ($E_{\bar{x}}$) and intra-ensemble standard deviation (E_{σ}) of each distribution represent an estimated mean value and internal variability of a given ensemble for a given epoch length at a given time (time is considered only for historical ensembles).

2.2.3. Degrees of Freedom

When considering time series, SSTA at a given time is highly correlated to several preceding/succeeding time steps leading to some predictability (e.g., McPhaden, 2003). However, to perform rigorous statistical analysis one must use independent values. One way to take into account the fact that time series are correlated to themselves

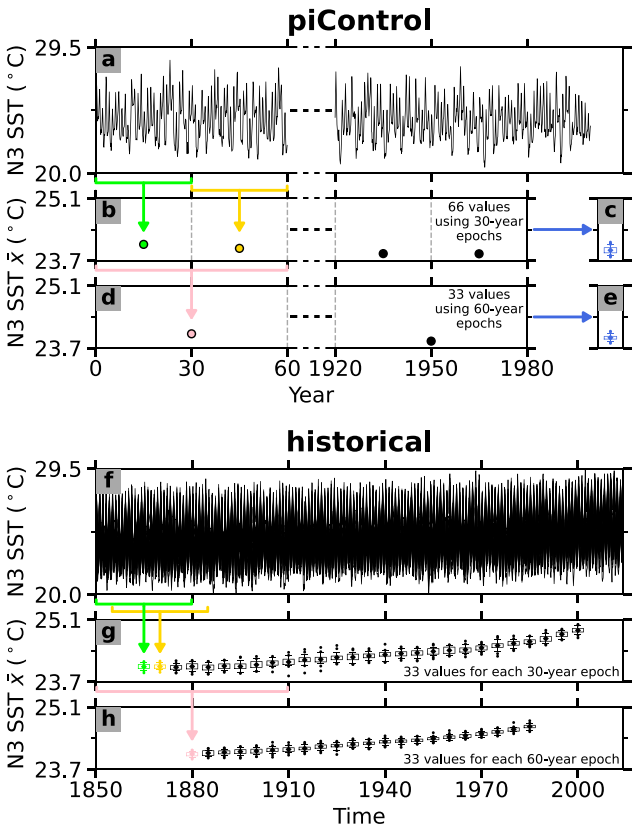

Figure 2. Graph describing how distributions are created from times series. Time series of Niño3 averaged sea surface temperature (N3 SST), from (a) a piControl run and (f) an ensemble of the historical run, as simulated by the IPSL-CM6A-LR model (Boucher et al., 2020). N3 SST mean (Equation 1) computed from (b) 30-year and (d) 60-year non-overlapping epochs to create the respective piControl distributions (c, e). N3 SST mean (Equation 1) computed from the historical ensemble using (g) 30-year and (h) 60-year epochs. The operation is repeated every 5 years to cover the entire historical run (e.g., 1850–1879, 1855–1884, 1860–1889, etc.). In the boxplots, whiskers extend to the 5th and 95th percentiles; boxes encompass the 25th and 75th percentiles; a diamond marks the mean; and dots indicate values that fall outside the whiskers.

(i.e., autocorrelation) is to reduce the number of time steps n to the number of effectively independent time steps (n^*). In statistics, this number is called "number of degrees of freedom", and can be estimated using:

$$n^* = \frac{n}{1 + \sum_{i=1}^{L} \rho_i^2} \tag{4}$$

where the autocorrelation function (ρ) is summed over the number of time steps (L) necessary to reach the first two sign changes (e.g., Atwood et al., 2017; Russon et al., 2014). Note that if the studied variable is stable through time (i.e., the autocorrelation function does not change), increasing the length of the time series by a factor α (e.g., $m = \alpha n$), will increase the number of degrees of freedom by the same factor ($m^* = \alpha n^*$).

Distributions of values from piControl runs are created using non-overlapping epochs to ensure the independence of each sample. Note that members of an ensemble are independent by construction.

2.2.4. Combinations

In Sections 3.1 and 3.3 the intra-ensemble standard deviation (E_σ) is computed using a given sample size (k) which is smaller or equal to the ensemble size (N). To do so, combinations (meaning that the order does not matter) of k distinct members of the ensemble are generated. The number of combinations used depends on the ensemble size and the sample size. If a large number of combinations are possible, 10,000 distinct combinations are randomly selected. The statistic is then averaged across combinations.

2.2.5. Standard Errors

Given a random sample $[x_1,\cdots, x_n]$ from a normal distribution $N(\mu, \sigma^2)$, the Standard Error (SE) of the sample mean ($SE_{\bar{x}}$; e.g., Chapter 4 p. 76 of von Storch & Zwiers, 1999), sample variance (SE_{σ^2}; e.g., Chapter 4 p. 77 of von Storch & Zwiers, 1999) and sample skewness (SE_{g_1}; e.g., Wright & Herrington, 2011) are:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{5}$$

$$SE_{\sigma^2} = \sigma^2 \sqrt{\frac{2}{n-1}} \tag{6}$$

$$SE_{g_1} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}} \tag{7}$$

where n is the number of independent samples (i.e., n^* for time series).

2.2.6. Confidence Intervals and Uncertainty of the Ensemble Mean

Using this random sample $[x_1,\cdots, x_n]$, the $p \times 100\%$ confidence intervals of the true (unknown) mean μ is (e.g., Chapter 5 p. 92 of von Storch & Zwiers, 1999):

$$\left(\bar{x} - Z\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z\frac{\sigma}{\sqrt{n}}\right) \tag{8}$$

where Z is the $0.5 + p/2$ quantile of the normal distribution and n is the number of independent samples (i.e., n^* for time series). In the paper the 95% confidence interval is used ($Z = 1.96$).

If we approximate the distribution of statistics computed on each member of an ensemble with a normal distribution (central limit theorem; e.g., Chapter 2 p. 35 of von Storch & Zwiers, 1999), we can define the absolute uncertainty of the ensemble mean (Δ) as the error on each side of the true (unknown) ensemble mean:

$$\Delta = Z \frac{E_\sigma}{\sqrt{N}} \tag{9}$$

where E_σ is the intra-ensemble standard deviation and N is the ensemble size.

It is sometimes useful to define the uncertainty relative to intra-ensemble mean ($E_{\overline{x}}$), hereafter "relative uncertainty" ($\Delta_r = 100\Delta/E_{\overline{x}}$). However, the relative uncertainty can become minuscule when $E_{\overline{x}} \gg \Delta$ (e.g., for N3 SST mean; not shown), or gigantic when $E_{\overline{x}} \ll \Delta$ (e.g., for N3 SSTA skewness; not shown). For simplicity, we use the absolute uncertainty (Δ) in almost all sections. The relative uncertainty (Δ_r) is only used in Section 3.4 in some cases. The main results of this paper are not altered if the relative uncertainty is used (not shown) and we verified that the uncertainties computed with Equation 9 are very similar to that computed using random sampling (see Text S3 and Figure S2 in Supporting Information S1).

3. Results

3.1. Influence of the Ensemble Size on the Uncertainty

In the literature, the uncertainty of the intra-ensemble mean (Δ) is usually computed with a random sampling and authors define one ensemble size for one given uncertainty (e.g., Maher et al., 2018; Milinski et al., 2020; J. Lee et al., 2021). Using Equation 9, one can analyze the relationship between ensemble size and uncertainty, as well as confronting our results with the theory: the uncertainty of ensemble mean should decrease with the square root of the ensemble size.

Figure 3 shows the ratio of the absolute uncertainty (Δ) computed with piControl and historical ensembles using combinations (see Section 2.2.4) of 10 to the maximum number of members (every 5 members) divided by Δ computed with the maximum number of members. Therefore, the horizontal axis represents the fraction of the ensemble size used for the computation. The results are presented for epoch lengths ranging from 30 to 150 years (15-year intervals) from the CMIP6-MME and 14 LEs with at least 15 members. We select larger LEs here compared to our initial threshold as we are creating synthetic ensembles of a smaller sizes and the minimum size of these synthetic ensembles is 10. There are a total of 188 curves (15 data sets x nine epoch lengths = 135 for the historical run, and 53 for the piControl run as the ensemble size decreases and falls below the 15 members threshold when the epoch length increases). All 15 data sets align almost perfectly on the theory (dashed black lines) for all three statistical moments computed with N3 PR and N3 SST. This means that even if we use very small samples, the theory can still be used. This is incredibly useful as, if one has an ensemble and wants to divide the uncertainty of the ensemble mean by 2, one immediately knows that the ensemble size must be multiplied by 4.

The only notable discrepancy comes from the piControl ensembles of N3 PRA variance computed with the MPI-ESM1-2-LR (blue cross markers in Figure 3c). This is due to the tendency of MPI-ESM1-2-LR to produce extremely rare but extremely large N3 PRA during El Niño events, resulting in a poorer convergence toward theoretical estimate for the smallest ensembles. In the 1000-year piControl simulation, anomalies of 5 mm.day^{-1} are reached during five events (equivalent to ∼9 standard deviations), including one reaching more than 9 mm.day^{-1} (more than 16 standard deviations). If these events are removed, this simulation falls back in the rank and follows the theory (see Figure S3 in Supporting Information S1). Note that such outliers inflicting a deviation from the theory are not found in the corresponding 50-member historical ensemble (i.e., 8250 years of simulation).

This analysis shows that, among 6 statistics computed with 30 different simulations, only one case deviated from the theory. Therefore, this approach is extremely robust.

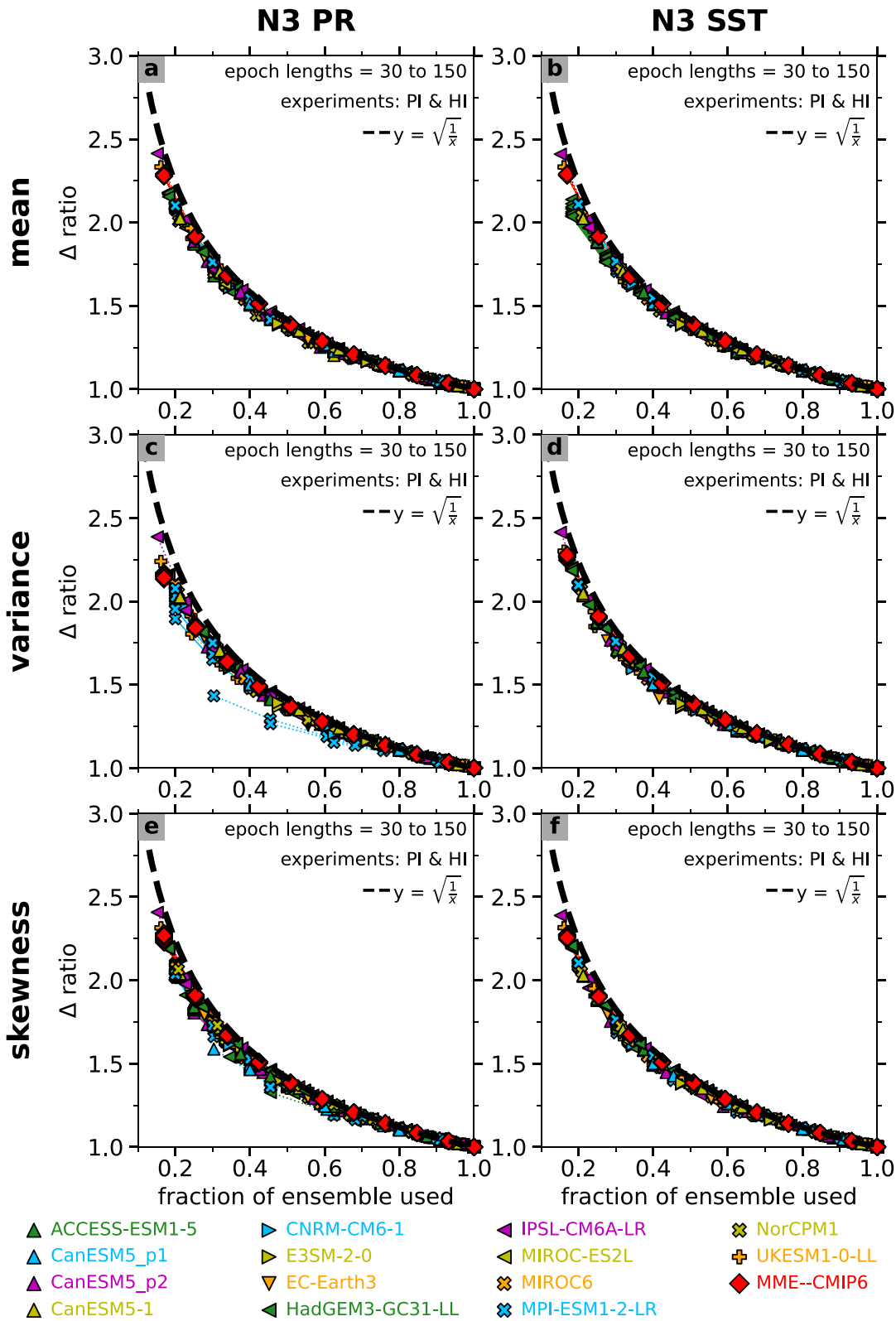

Figure 3. Dependence of the uncertainty of the ensemble mean (Δ; Equation 9) on the fraction of the ensemble used. Uncertainty computed for Niño3 averaged precipitation (a, c, e) and Niño3 averaged sea surface temperature (b, d, f) mean (a, b), variance (c, d) and skewness (e, f). The dashed black line in each panel represents the theoretical improvement of the uncertainty with the square root of the fraction of the ensemble used. The uncertainty of the ensemble mean is computed using all epoch lengths and all epochs of the piControl (dotted lines) and historical (solid lines) runs from 14 LEs with at least 15 members and the CMIP6-MME.

3.2. Influence of the Epoch Length on the Uncertainty

For simple diagnostics (like the first three statistical moments), one can use Equations 5–7 with σ and n respectively equal to the standard deviation of the time series and the number of independent time steps (n^*), and notice that, in theory, the error of these statistics decreases with the square root of the number of independent time steps.

Now, does it mean the intra-ensemble standard deviation (E_σ) decreases at the same rate when the epoch length is increased? Figure 4 shows the ratio of the uncertainty of the ensemble mean (Δ) computed with historical ensembles using epoch lengths of 30–150 years (15-year intervals) divided by Δ computed with 150-year epochs. Epoch lengths (i.e., time steps) are used instead of independent number of time steps (degrees of freedom; see Section 2.2.3) as the latter is proportional to the number of time steps: if T time steps are independent in a 150-year epoch, ∼T/2 are independent in a 75-year epoch (not shown). The results are presented for all 26 LEs and the CMIP6-MME, using the maximum number of members of each ensemble (27 curves in each panel). Although the magnitude of the uncertainty reduction is more model dependent than for the influence of the ensemble size (Section 3.1), most ensembles show a change of Δ that is broadly consistent with the theory (dashed black lines) for all three statistical moments computed with N3 PR and N3 SST.

However, for N3 SST mean (Figure 3b) and N3 PRA skewness (Figure 3e), several ensembles are clearly departing from the theory. For these ensembles and diagnostics, the intra-ensemble standard deviation (E_σ) is not increasing as fast as expected (or even decreases), with decreasing epoch length. The exact reason is beyond the scope of this paper but two simple reasons may explain this result: (a) Equations 5–7 are valid when the sample is drawn from a normal distribution but N3 PRA and N3 SSTA (to a smaller extent) distributions are skewed (Figures 1j and 1l); and (b) a small sized LE can randomly deviate from the theory (see Text S4 and Figure S4 in Supporting Information S1).

The behavior of the CMIP6-MME is also notable: varying the epoch length has no influence on the uncertainty. This is due to the fact that increasing the epoch length only attenuates the internal variability within each model, it does not reduce the inter-model differences (inter-model differences shown by the red boxplots in Figures 1b, 1d, 1f, 1h, 1j, and 1l). So, if one wants to detect a change in a statistical value (e.g., related to climate change) using the CMIP6-MME, increasing the epoch length will not reduce the uncertainty. One may detect a change only if it is large enough between the considered epochs.

While the uncertainty should similarly increase with decreasing epoch length in the piControl run, it is not easy to prove it due to the methodology used to create the distributions (see Section 2.2.2). Indeed, with the piControl run increasing the epoch length implies a smaller number of samples, reducing our ability to robustly compute the standard deviation of the distribution (E_σ). Despite this methodological issue, with a long piControl run, the uncertainty of the ensemble mean does follow the theory (see Text S5 and Figure S5 in Supporting Information S1).

Thus, both ensemble size and epoch length can be used to improve the uncertainty of the ensemble mean to obtain a more robust evaluation of the climate models. However, decreases in uncertainty with increasing the ensemble size almost perfectly follow expectations from theory, while increasing the epoch length may not have the desired influence if time series are not relatively constant or for diagnostics more complex than the first three statistical moments.

3.3. Uncertainty in piControl Versus Historical Runs

We compare now the uncertainty of the ensemble mean (Δ) computed from the 26 historical LEs and the CMIP6-MME with the corresponding piControl runs (Figure 5), using combinations of k members (see Section 2.2.4), k being the minimum sample size between the historical and piControl distributions. Here, we only use 30-year epochs as some piControl runs are only 300-year long, that is, 10 non-overlapping epochs, which is already a relatively small sample size to compute a standard deviation (we verified that the relationship is similar with other epoch length; not shown). The CMIP6-MME is not included for the N3 SST mean (Figure 5b) as the uncertainty is ∼100% larger than the largest uncertainty computed with LEs and would spuriously increase the correlations (not shown). This is linked to the fact that the difference from one model to another (the mean state bias of the models; red boxplot in Figure 1d) is much larger than the difference between a member of a given model to another

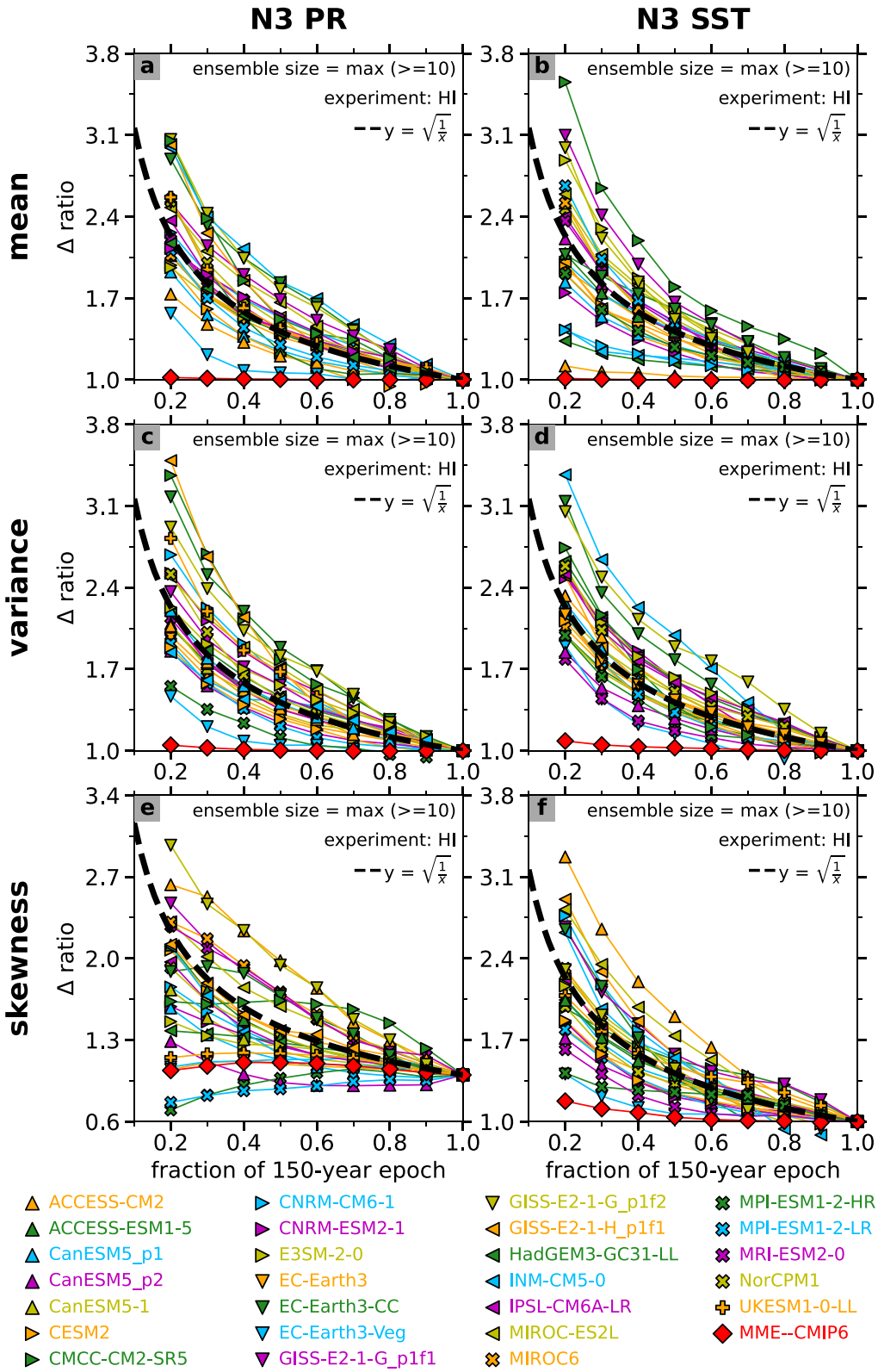

Figure 4. Dependence of the uncertainty of the ensemble mean (Δ; Equation 9) on the fraction of 150-year used for the computation. Uncertainty computed for Niño3 averaged precipitation (a, c, e) and Niño3 averaged sea surface temperature (b, d, f) mean (a, b), variance (c, d) and skewness (e, f). The dashed black line in each panel represents the theoretical improvement of the uncertainty with the square root of the fraction of 150-year used. Uncertainty computed using all epochs of the historical run from CMIP6-MME and all 26 LEs (using the maximum ensemble size). Note that panel e does not have the same vertical range as the other panels.

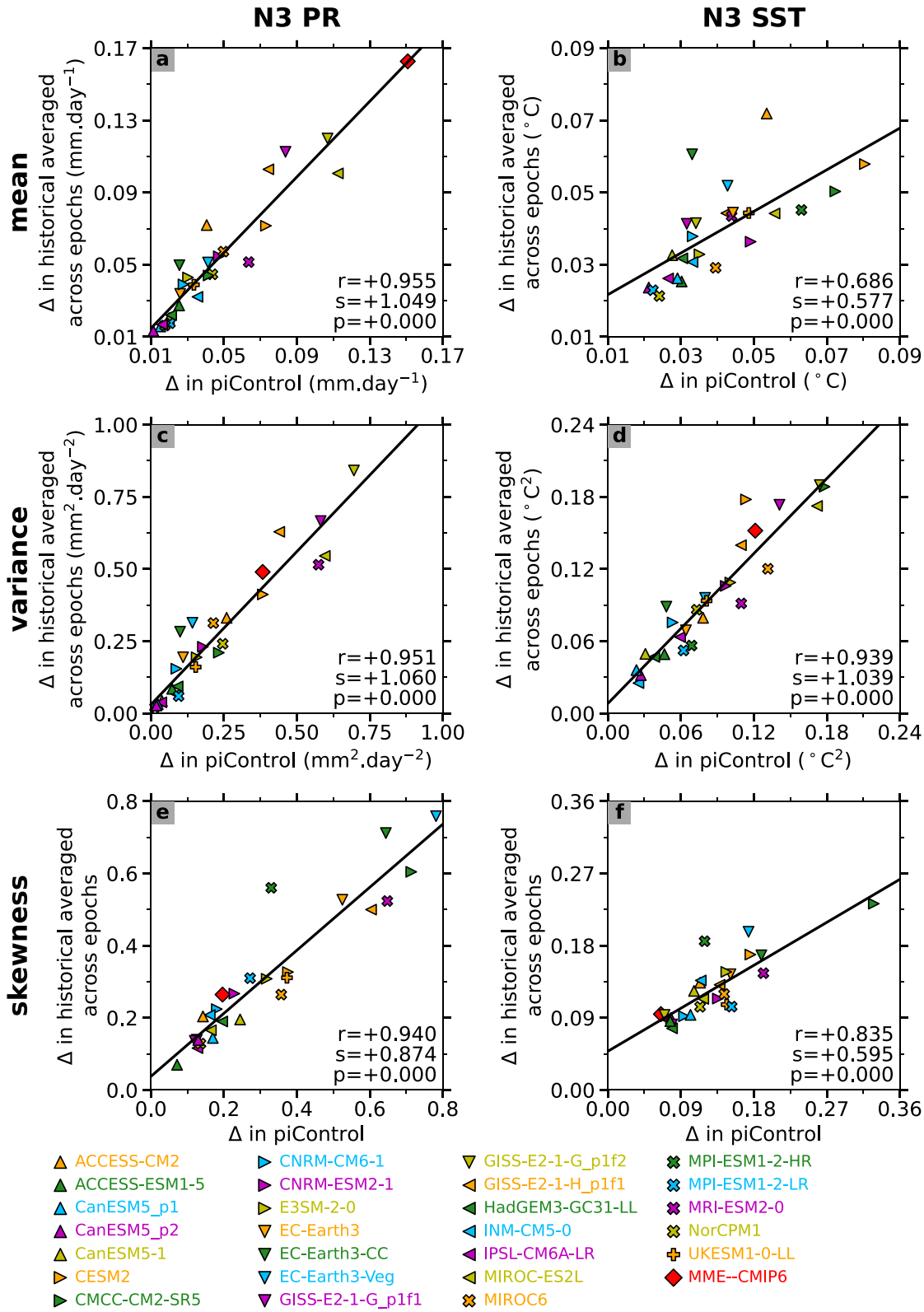

Figure 5.

member of the same model (i.e., the mean state modulation by the internal variability; purple boxplot in Figure 1d).

This analysis reveals that four of the six diagnostics (Figures 5a, 5c, 5d, and 5e) produce an almost perfect match between historical and piControl runs (correlation >0.9, slope ∼1, intercept ∼0). The relationship is not as good in the other two diagnostics (correlation ∼0.7, slope ∼0.6, intercept >0; Figures 5b and 5f), with better uncertainties (i.e., smaller Δ) in the historical run compared to the piControl run when the uncertainty value is large. Overall, there is a good correspondence between the uncertainty of the ensemble means computed with piControl and historical runs. This means that even though external forcing has affected the climate system, the climate is not radically different (ENSO related statistics computed here don't change much), and the internal variability of the climate is not radically different (the uncertainties of the ENSO related statistics computed here don't change much). This implies that the control simulation can be used when the historical ensemble is small, or to estimate the size of the historical ensemble before computing it. This is useful for modelers because multiple control runs may be performed during the model development or tuning process, well-before historical runs are performed.

3.4. Estimating the Ensemble Size

There are many papers in the literature proposing a minimum number of members, often termed the RES, that should be computed for a particular application such as ENSO (e.g., Maher et al., 2018; Milinski et al., 2020; J. Lee et al., 2021). Here, we propose to use the theory instead to estimate the RES, by rearranging Equation 9:

$$\mathrm{RES} = \left(Z\frac{E_\sigma}{\Delta}\right)^2 \tag{10}$$

This way, one can easily estimate the RES given an absolute (Δ) or relative ($\Delta_r = 100\Delta/E_{\bar{x}}$) uncertainty. The main advantage of computing the RES using Equation 10 is that it is not limited by the size of the existing ensemble (which is one limitation of computing the RES using random sampling). Note that both methods lead to equivalent results (see Text S6 and Figure S5 in Supporting Information S1). Another advantage is that, if a piControl simulation is available, one can use Equation 10 to estimate the RES for a particular application, before the ensemble is generated.

We propose in this section to apply this method to all 59 CMIP6 ensembles using piControl runs to explore possible ensemble sizes for all models. This approach has some limitations as the mean statistic and the internal variability may change between piControl and historical runs. We also limit it to 60 members even if in some cases more members would be needed. We decided to cap the number of members as we aim here to describe methodologies and order of magnitudes, not to provide exact numbers. In addition, this cap is already larger than any LE computed for past CMIP exercises.

There are several ways in which this proposed formula may be utilized. Firstly, one can estimate the RES to reach a given uncertainty. Here we estimate the RES needed to reach an absolute uncertainty (Δ) of 0.05°C and 0.1 for N3 SST mean and skewness respectively (Figures 6b and 6f; gold), a relative uncertainty (Δ_r) of 5% for N3 PR mean (Figure 6a; gold) and of 20% for N3 PR variance and skewness, as well as N3 SST variance (Figures 6c, 6d and, 6e; gold). To reach these uncertainties, the IPSL-CM6A-LR ensemble (purple triangles) requires less than 20 members. On average across CMIP6 ensembles (boxplot), less than 30 members are required, while focusing on individual models three models require ensembles with more than 60 members for N3 PR variance and N3 SST skewness. It is also interesting to note that the RES can be three times larger for N3 PR variance compared to N3 SST variance to reach the same relative uncertainty (20%), meaning that the internal variability of N3 PR variance is larger relative to that of N3 SST variance. This is likely linked to the fact that precipitation is more nonlinear (e.g., Frauen et al., 2014; Garfinkel et al., 2018; Sun et al., 2016), implying stronger interdecadal modulation of its variance.

Figure 5. Uncertainty of the ensemble mean (Δ; Equation 9) computed from historical versus the piControl runs. Uncertainty computed using 30-year epochs for Niño3 averaged precipitation (a, c, e) and Niño3 averaged sea surface temperature (b, d, f) mean state (a, b), variance (c, d) and skewness (e, f). Uncertainty computed in the 26 LEs and the CMIP6-MME using the minimum sample size of historical and the piControl runs. For the historical run, the uncertainty is computed for all epochs and averaged. The solid black line in each panel represents the linear regression. The corresponding correlation (*r*), regression slope (*s*) and p-value (*p*) are indicated at the bottom of each panel.

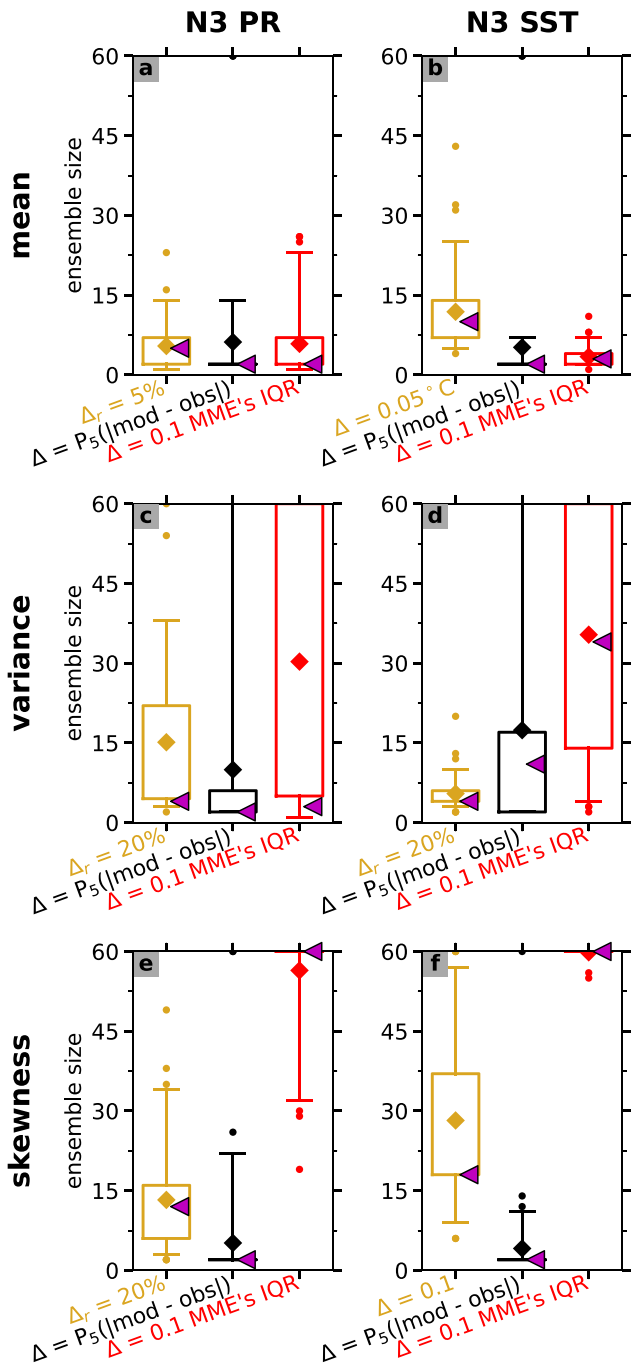

Figure 6. Ensemble sizes required to limit the uncertainty to a desired value (Equation 10). Required ensemble size (RES) computed with the 59 piControl distributions to reach a given uncertainty (gold), to know the sign of the model bias at the 95% confidence level ($\Delta = P_5|E_{\bar{x}} - obs|$; black) and to limit the overlap of the confidence interval of each model ($\Delta = 0.1$CMIP6′sIQR; red). RES computed for Niño3 averaged precipitation (a, c, e) and Niño3 averaged sea surface temperature (b, d, f) mean (a, b), variance (c, d) and skewness (e, f). Purple triangles represent the RES for IPSL-CM6A-LR. Boxplots represent the distribution of values computed using all model intercomparison project phase 6 ensembles: whiskers extend to the 5th and 95th percentiles; boxes encompass the 25th and 75th percentiles; a diamond marks the mean; and dots indicate values that fall outside the whiskers.

Secondly, one may want to know the sign of the ensemble's bias and set the absolute uncertainty to a confidence interval on the absolute difference between ensemble mean and observational data set (for the 95% confidence interval, $\Delta = P_5|E_{\bar{x}} - obs|$; Figure 6; black). Knowing the sign of the bias can be usually achieved with less than 20 members for all CMIP6 ensembles (e.g., 11 is the maximum RES needed for the IPSL-CM6A-LR ensemble). In some cases, the RES can be very high because the model bias is extremely small (this was also the case in J. Lee et al. (2021)). A second criteria could be introduced to avoid this issue, for example, limiting the desired uncertainty with a fraction of the observed value (e.g., $\Delta = max(P_5|E_{\bar{x}} - obs|, 0.05\ obs)$).

Finally, one can desire a robust ranking of CMIP6 ensembles, implying to limit the overlap of the confidence interval of each model. This can be done by setting the absolute uncertainty to a fraction of the CMIP6 distribution ($\Delta = 0.1$CMIP6′sIQR; Figure 6; red). In this case, CMIP6 ensembles (boxplot) can be correctly ranked only for N3 PR and N3 SST means, for which no ensemble needs to be larger than 27. For the other four diagnostics (N3 PR and N3 SST variances and skewness) the desired uncertainty is largely out of reach (i.e., ∼30% of ensembles do not reach it within 60 members for N3 PR and N3 SST variances, and ∼85% for N3 PR and N3 SST skewness). Note that the desired uncertainty specified is quite loose, in that even if it is reached, ranking of models would be difficult. For instance, 30 ensembles are found within the IQR and each of their ensemble means would be within a range equivalent to 0.2xIQR (0.1 IQR on each side of the mean), implying an important overlap between the uncertainty of each ensemble. According to our results, it would be hard to provide a robust ranking of CMIP6 ensembles for N3 PR and N3 SST variances and virtually impossible to do it for N3 PR and N3 SST skewness.

4. Conclusions

We analyzed the first three statistical moments (mean, variance, and skewness) of N3 PR and N3 SST computed from all available CMIP6 piControl and historical ensembles (26 large ensembles and the CMIP6-MME consisting of the first ensemble member from each of 59 different model configurations) to better describe how ensemble means are influenced by ensemble size and the length of the epoch used to compute the statistic. The key results are the following:

- The uncertainty of the intra-ensemble mean (Δ) decreases with the square root of the ensemble size, in accordance with theory. Thus, if one has an ensemble mean with an uncertainty Δ, and wishes to reduce that uncertainty to Δ/2, the ensemble size must be quadrupled.
- The epoch length generally has a similar effect on Δ. However, this does not apply to a MME, and there are more inter-model differences in this relationship–possibly linked to the non-normality and/or multi-decadal modulation of some model distributions, and the relatively small sizes of some of the available model ensembles.
- There is a good correspondence between Δ computed from a historical LE and that computed from the same model's piControl. This implies that one can use a piControl run to estimate the Δ for a given historical ensemble, or to estimate, or to estimate how many historical ensemble members must be generated to obtain a given Δ,.
- With our piControl-based method, one can simply estimate the ensemble size required to fit one's purpose, regardless of the ensemble size already computed. This contrasts with randomly sampling an existing ensemble

(as in Milinski et al., 2020), where one can estimate the RES only if it is smaller than the one already computed.

The first two key results are expected according to statistical theory, and here we confirmed them using relatively small (but practically relevant) samples derived from climate data (10–50 members; 30–150 years of simulation). Sample sizes as small as 10 (smallest size tested) are large enough to confirm the expected relationship between Δ and the ensemble size. Only cases with extreme outliers are shown to deviate from the theory (e.g., N3 PRA variance from MPI-ESM1-2-LR's piControl). Larger sample sizes (30 or more) are required to confirm the relationship between Δ and the number of time steps. For the skewness, this relationship critically depends on how the data are distributed (Wright & Herrington, 2011), so one should not expect the same decrease of Δ with increasing number of time steps as for the mean or the variance. The third key result relates to the simulated internal variability in piControl and historical runs. In some cases, internal variability is larger in historical runs, particularly for large internal variability or small ensemble sizes. Further investigation is required to understand such discrepancy.

The method that we propose to estimate the RES complements the random sampling methods of Milinski et al. (2020) and J. Lee et al. (2021), but at a much smaller computation cost (no random sampling). Specifically, the resampling is replaced by a mathematical formula to compute Δ (Equation 9) or by the ensemble size required to achieve a given Δ (Equation 10). Our equations can be used by any model user to fit their own purpose, and show how an existing computation done with a given ensemble size and epoch length can be used to estimate Δ for other ensemble sizes or epoch lengths. Although we used some simple diagnostics (mean, variance, and skewness of area-averaged SST and precipitation) to demonstrate applications of the theory, Equations 9 and 10 can potentially be used more broadly for other diagnostics and variables, if they are sufficiently normally distributed.

As an example, we propose that this framework could be applied to analyze Tropical Pacific Decadal Variability (TPDV; Power et al., 2021) and related mechanisms (Capotondi et al., 2023) in climate models. To detect and attribute a decadal change to external radiative forcings, one requires an ensemble large enough so that Δ is smaller than the radiatively forced changes. Similarly, when analyzing the influence of natural forcings in long paleoclimate simulations (e.g., orbit eccentricity and tilt, CO_2 concentration; Yun et al., 2023) or comparing past and future climates (e.g., Brown et al., 2020) knowing the sampling uncertainty for the target climate phenomenon within the available observations and ensemble simulations is crucial to obtain robust results. Finally, intermediate models (e.g., Linear Inverse Models, Penland & Sardeshmukh, 1995; ZC model, Zebiak & Cane, 1987) are often used to obtain long simulations (e.g., 100,000 years in Ramesh & Cane, 2019) at a low cost compared to coupled general circulation models. But how long must these runs be, for a given application? In these cases, one can use Equation 9 to estimate if the simulation length and ensemble size are sufficient, and/or Equation 10 to estimate how many additional years (or members) must be computed to reach a desired accuracy.

Data Availability Statement

CMIP6 data can be accessed at https://esgf-node.llnl.gov/projects/esgf-llnl/. Global Precipitation Climatology Project Monthly Analysis Product version 2.3 (GPCPv2.3; Adler et al., 2003) and NOAA Optimum Interpolation Sea Surface Temperature version 2 (OISSTv2; Reynolds et al., 2002) data products are provided by NOAA PSL, Boulder, Colorado, USA, and available from their website at https://psl.noaa.gov/. Data sets were analyzed using the CLIVAR ENSO metrics package (Planton et al., 2021; https://github.com/CLIVAR-PRP/ENSO_metrics), executed via the PCMDI Metrics Package framework (Lee et al., 2024; https://github.com/PCMDI/pcmdi_metrics). The output and processing scripts used for the paper (Planton & Lee, 2024) are available at https://zenodo.org/doi/10.5281/zenodo.11512024. Additional figures are available at: https://yyplanton.github.io/estimating_uncertainties_enso/.

Acknowledgments
The authors thank the four anonymous reviewers for their constructive comments and suggestions. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for organizing CMIP. We thank all the international climate modeling groups for their tireless development efforts, and for generously producing and publishing these coordinated, standardized, and quality-controlled simulations. We thank the Earth System Grid Federation (ESGF) for archiving the simulations and improving access, and are grateful to the multiple funding agencies who support CMIP and ESGF. The U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support for CMIP, and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. Author Planton held a National Research Council Research Associateship at NOAA/PMEL when he started this work. He is now supported by the Australian Government's National Environmental Science Program (NESP2) Climate Systems Hub. Author McGregor was supported by NESP2 Climate Systems Hub and the Australian Research Council (Grant FT160100162 and DP200102329). Work of LLNL-affiliated authors was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, with their efforts supported by the Regional and Global Model Analysis (RGMA) program of the United States Department of Energy's Office of Science. We also acknowledge the support of the ARISE ANR (Agence Nationale pour la Recherche, France) project (ANR-18-CE01-0012). PMEL contribution no. 5394. Open access publishing facilitated by Monash University, as part of the Wiley - Monash University agreement via the Council of Australian University Librarians.

References

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

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

Anderson, W., Seager, R., Baethgen, W., & Cane, M. (2018). Trans-Pacific ENSO teleconnections pose a correlated risk to agriculture. *Agricultural and Forest Meteorology*, *262*, 298–309. https://doi.org/10.1016/j.agrformet.2018.07.023

Atwood, A. R., Battisti, D. S., Wittenberg, A. T., Roberts, W. H. G., & 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, E., 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

Bertrand, A., Lengaigne, M., Takahashi, K., Avadí, A., Poulain, F., & Harrod, C. (2020). El Niño Southern Oscillation (ENSO) effects on fisheries and aquaculture. In *FAO fisheries and aquaculture technical paper No. 660*. FAO. https://doi.org/10.4060/ca8348en

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

Brown, J. R., Brierley, C. M., An, S.-I., Guarino, M.-V., Stevenson, S., Williams, C. J. R., et al. (2020). Comparison of past and future simulations of ENSO in CMIP5/PMIP3 and CMIP6/PMIP4 models. *Climate of the Past*, *16*(5), 1777–1805. https://doi.org/10.5194/cp-16-1777-2020

Cai, W., Ng, B., Wang, G., Santoso, A., Wu, L., & Yang, K. (2022). Increased ENSO sea surface temperature variability under four IPCC emission scenarios. *Nature Climate Change*, *12*(3), 228–231. https://doi.org/10.1038/s41558-022-01282-z

Capotondi, A., McGregor, S., McPhaden, M. J., Cravatte, S., Holbrook, N. J., Imada, Y., et al. (2023). Mechanisms of tropical Pacific decadal variability. *Nature Reviews Earth & Environment*, *4*(11), 754–769. https://doi.org/10.1038/s43017-023-00486-x

Cashin, P., Mohaddes, K., & Raissi, M. (2017). Fair weather or foul? The macroeconomic effects of El Niño. *Journal of International Economics*, *106*, 37–54. https://doi.org/10.1016/j.jinteco.2017.01.010

Chen, Y., Morton, D. C., Andela, N., van der Werf, G. R., & Randerson, J. T. (2017). A pan-tropical cascade of fire driven by El Niño/Southern Oscillation. *Nature Climate Change*, *7*(12), 906–911. https://doi.org/10.1038/s41558-017-0014-8

Cramér, H. (1946). *Mathematical methods of statistics (PMS-9)* (Vol. 9). Princeton University Press. https://doi.org/10.1515/9781400883868

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/10.1038/s41558-020-0731-2

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*, 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 [Dataset]. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Frauen, C., Dommenget, D., Tyrrell, N., Rezny, M., & Wales, S. (2014). Analysis of the nonlinearity of El Niño-Southern oscillation teleconnections. *Journal of Climate*, *27*(16), 6225–6244. https://doi.org/10.1175/JCLI-D-13-00757.1

Garfinkel, C. I., Gordon, A., Oman, L. D., Li, F., Davis, S., & Pawson, S. (2018). Nonlinear response of tropical lower-stratospheric temperature and water vapor to ENSO. *Atmospheric Chemistry and Physics*, *18*(7), 4597–4615. https://doi.org/10.5194/acp-18-4597-2018

Goddard, L., & Gershunov, A. (2020). Impact of El Niño on weather and climate extremes. In M. J. McPhaden, A. Santoso, & W. Cai (Eds.), *El Niño southern oscillation in a changing climate, geophysical monograph* (Vol. 253, pp. 361–375). American Geophysical Union. https://doi.org/10.1002/9781119548164.ch16

Jin, F.-F., Chen, H.-C., Zhao, S., Hayashi, M., Karamperidou, C., Stuecker, M. F., et al. (2020). Simple ENSO models. In M. J. McPhaden, A. Santoso, & W. Cai (Eds.), *El Niño southern oscillation in a changing climate, geophysical monograph* (Vol. 253, pp. 119–151). American Geophysical Union. https://doi.org/10.1002/9781119548164.ch6

Kestin, T. S., Karoly, D. J., Yano, J.-I., & Rayner, N. A. (1998). Time–frequency variability of ENSO and stochastic simulations. *Journal of Climate*, *11*(9), 2258–2272. https://doi.org/10.1175/1520-0442(1998)011<2258:TFVOEA>2.0.CO;2

Lee, J., Gleckler, P. J., Ahn, M.-S., Ordonez, A., Ullrich, P. A., Sperber, K. R., et al. (2024). Objective evaluation of Earth system models: PCMDI Metrics Package (PMP) version 3. *Geoscientific Model Development*, *17*(9), 3919–3948. https://doi.org/10.5194/gmd-17-3919-2024

Lee, J., Planton, Y. Y., Gleckler, P. J., Sperber, K. R., Guilyardi, E., Wittenberg, A. T., et al. (2021a). Robust evaluation of ENSO in climate models: How many ensemble members are needed? *Geophysical Research Letters*, *48*(20), e2021GL095041. https://doi.org/10.1029/2021GL095041

Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J. P., et al. (2021b). Future global climate: Scenario-based projections and near-term information. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, et al. (Eds.), *Climate change 2021: The physical science basis. Contribution of working group I to the Sixth assessment Report of the intergovernmental panel on climate change*. Cambridge University Press. https://doi.org/10.1017/9781009157896.006

Maher, N., Matei, D., Milinski, S., & Marotzke, J. (2018). ENSO change in climate projections: Forced response or internal variability? *Geophysical Research Letters*, *45*(20), 11390–11398. https://doi.org/10.1029/2018GL079764

McPhaden, M. J. (2003). Tropical Pacific Ocean heat content variations and ENSO persistence barriers. *Geophysical Research Letters*, *30*(9), 1480. https://doi.org/10.1029/2003GL016872

McPhaden, M. J., Santoso, A., & Cai, W. (Eds.) (2020). *El Niño Southern oscillation in a changing climate, Geophysical monograph* (Vol. 253). American Geophysical Union. https://doi.org/10.1002/9781119548164

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., & Stouffer, R. J. (2000). The Coupled Model Intercomparison Project (CMIP). *Bulletin of the American Meteorological Society*, *81*(2), 313–318. https://doi.org/10.1175/1520-0477(2000)080<0305:MROTEA>2.3.CO;2

Meehl, G. A., Covey, C., Delworth, D., Latif, M., McAvaney, B., Mitchell, J. F. B., 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*(4), 885–901. https://doi.org/10.5194/esd-11-885-2020

Ng, B., Cai, W., Cowan, T., & Bi, D. (2021). Impacts of low-frequency internal climate variability and greenhouse warming on El Niño-southern oscillation. *Journal of Climate*, *34*(6), 2205–2218. https://doi.org/10.1175/JCLI-D-20-0232.1

Penland, C., & Sardeshmukh, P. D. (1995). The optimal growth of tropical Sea Surface temperature anomalies. *Journal of Climate*, *8*(8), 1999–2024. https://doi.org/10.1175/1520-0442(1995)008<1999:TOGOTS>2.0.CO;2

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

Planton, Y. Y., & Lee, J. (2024). Data and codes for the paper "Estimating uncertainty in simulated ENSO statistics [Dataset]. *Zenodo*. https://doi.org/10.5281/zenodo.11512024

Power, S., Lengaigne, M., Capotondi, A., Khodri, M., Vialard, J., Jebri, B., et al. (2021). Decadal climate variability in the tropical Pacific: Characteristics, causes, predictability, and prospects. *Science*, *374*(6563), eaay9165. https://doi.org/10.1126/science.aay9165

Ramesh, N., & Cane, M. A. (2019). The predictability of tropical Pacific decadal variability: Insights from attractor reconstruction. *Journal of the Atmospheric Sciences*, *76*(3), 801–819. https://doi.org/10.1175/JAS-D-18-0114.1

Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., & Wang, W. (2002). An improved in situ and satellite SST analysis for climate [Dataset]. *Journal of Climate*, *15*(13), 1609–1625. https://doi.org/10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2

Russon, T., Tudhope, A. W., Hegerl, G. C., Schurer, A., & Collins, M. (2014). Assessing the significance of changes in ENSO amplitude using variance metrics. *Journal of Climate*, *27*(13), 4911–4922. https://doi.org/10.1175/JCLI-D-13-00077.1

Sun, Y., Wang, F., & Sun, D.-Z. (2016). Weak ENSO asymmetry due to weak nonlinear air-sea interaction in CMIP5 climate models. *Advances in Atmospheric Sciences*, *33*(3), 352–364. https://doi.org/10.1007/s00376-015-5018-6

Taschetto, A. S., Ummenhofer, C. C., Stuecker, M. F., Dommenget, D., Ashok, K., Rodrigues, R. R., & Yeh, S. W. (2020). ENSO atmospheric teleconnections. In M. J. McPhaden, A. Santoso, & W. Cai (Eds.), *El Niño Southern oscillation in a changing climate, geophysical monograph* (Vol. 253, pp. 361–375). American Geophysical Union. https://doi.org/10.1002/9781119548164.ch14

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Thompson, D. W. J., Barnes, E. A., Deser, C., Foust, W. E., & Phillips, A. S. (2015). Quantifying the role of internal climate variability in future climate trends. *Journal of Climate*, *28*(16), 6443–6456. https://doi.org/10.1175/JCLI-D-14-00830.1

von Storch, H., & Zwiers, F. W. (1999). *Statistical analysis in climate Research*. Cambridge University Press. https://doi.org/10.1017/CBO9780511612336

Wittenberg, A. T. (2009). Are historical records sufficient to constrain ENSO simulations? *Geophysical Research Letters*, *36*(12), L12702. https://doi.org/10.1029/2009GL038710

Wright, D. B., & Herrington, J. A. (2011). Problematic standard errors and confidence intervals for skewness and kurtosis. *Behavior Research Methods*, *43*(1), 8–17. https://doi.org/10.3758/s13428-010-0044-x

Yun, K.-S., Lee, J.-Y., Timmermann, A., Stein, K., Stuecker, M. F., Fyfe, J. C., & Chung, E.-S. (2021). Increasing ENSO–rainfall variability due to changes in future tropical temperature–rainfall relationship. *Communications Earth & Environment*, *2*(1), 43. https://doi.org/10.1038/s43247-021-00108-8

Yun, K.-S., Timmermann, A., Lee, S.-S., Willeit, M., Ganopolski, A., & Jadhav, J. (2023). A transient Coupled General Circulation Model (CGCM) simulation of the past 3 million years. *Climate of the Past*, *19*(10), 1951–1974. https://doi.org/10.5194/cp-19-1951-2023

Zebiak, S. E., & Cane, M. A. (1987). A model El Niño–Southern oscillation. *Monthly Weather Review*, *115*(10), 2262–2278. https://doi.org/10.1175/1520-0493(1987)115<2262:AMENO>2.0.CO;2

Zheng, X.-T., Hui, C., & Yeh, S.-W. (2018). Response of ENSO amplitude to global warming in CESM large ensemble: Uncertainty due to internal variability. *Climate Dynamics*, *50*(11), 4019–4035. https://doi.org/10.1007/s00382-017-3859-7

References From the Supporting Information

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*, *14*(11), 7073–7116. https://doi.org/10.5194/gmd-14-7073-2021

Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., Russell, G. L., et al. (2020). GISS-E2.1: Configurations and climatology. *Journal of Advances in Modeling Earth Systems*, *12*(8), e2019MS002025. https://doi.org/10.1029/2019MS002025

Swart, N. C., Cole, J. N. S., 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