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, Victoria, Australia

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

[3]Lawrence Livermore National Laboratory, Livermore, California, USA

[4]NOAA Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, 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, Victoria, Australia

Corresponding author: Yann Y. Planton (yann.planton@monash.edu)

Key Points:

- Large ensembles of historical runs and long control runs from climate models are analyzed to study the uncertainty of the ensemble mean
- The uncertainty of the ensemble mean decreases with the square root of the ensemble size or the epoch length used to perform the calculation
- A simple equation yields an estimate of the ensemble mean uncertainty or the ensemble size needed to limit the uncertainty to a given value

Abstract

The use of large ensembles of model simulations is growing due to the need to minimize the influence of internal variability in evaluation of climate models and the detection of climate change induced trends. Yet, exactly how many ensemble members are required to effectively separate internal variability from climate change varies from model to model and metric to metric. Here we analyze the first three statistical moments (i.e., mean, variance and skewness) of detrended precipitation and sea surface temperature (interannual anomalies for variance and skewness) in the eastern equatorial Pacific from observations and ensembles of Coupled Model Intercomparison Project Phase 6 (CMIP6) climate simulations. We then develop/assess the equations, based around established statistical theory, for estimating the required ensemble size for a user defined uncertainty range. Our results show that — as predicted by statistical theory — the uncertainties in ensemble means of these statistics decreases with the square root of the time series length and/or ensemble size. Further to this, as the uncertainties of these ensemble-mean statistics are generally similar when computed using pre-Industrial control runs versus historical runs, the pre-industrial runs can sometimes be used to estimate: i) the number of realizations and years needed for a historical ensemble to adequately characterize a given statistic; or ii) the expected uncertainty of statistics computed from an existing historical simulation or ensemble, if a large ensemble is not available.

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 via consideration of the substantial high-performance computing resources.

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). It is 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 to 3°C.

With climate models being primary tools for improving our understanding of Earth's past, present and future climate, evaluation of ENSO in climate models has garnered substantial interest. Knowing how well climate models represent key aspects of the historical climate 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; Meehl et al., 2000, 2007; Taylor et al., 2012; Eyring et al., 2016) has enabled the benchmarking of model's 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 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 future ENSO changes with global warming (e.g., Wittenberg, 2009; Maher et al., 2018; Zheng et al., 2018; Ng et al., 2021) as well as robustly evaluating model performance (J. Lee et al., 2021). The use of model ensembles (multiple simulations with each model configuration) is an established approach to identify the impact of internal variability on model characteristics and projections (e.g., Deser et al., 2020).

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 ($\bar{x}$), variance (σ^2) and skewness (g_1) can be defined as follows (e.g., Cramér, 1946):

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

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

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

(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 precipitation (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 multi-model ensemble (MME) to that of the single-model initial condition ensemble (made of 33 Historical simulations of IPSL-CM6A-LR model; green boxplots), it is evident that initial conditions have 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 (Figure 1h,l).

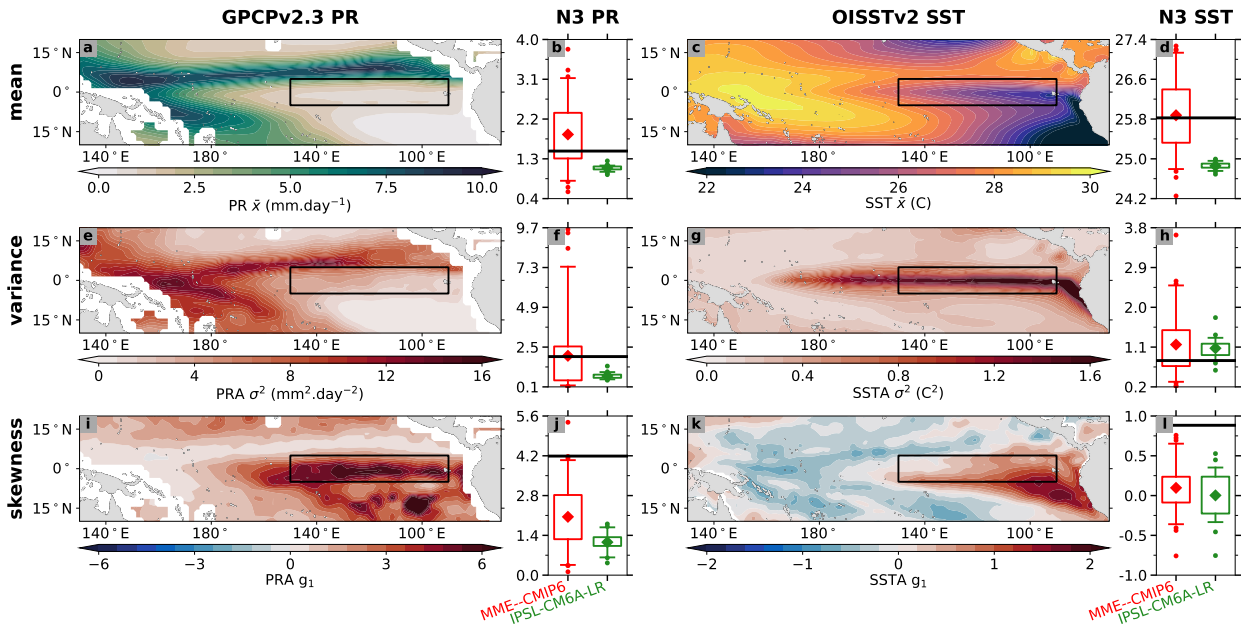


Figure 1. Statistical moments computed with observed and modeled PR and SST. Maps of observed PR (a,e,i; left column) and SST (c,g,k; right column) over the tropical Pacific Ocean, alongside Niño3 averaged (black rectangle) modeled (boxplots) and observed (black line) PR (b,f,j) and SST (d,h,l). Statistical 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 datasets. Boxplots represent the distributions of statistics computed from a multi-model ensemble (MME; 59 CMIP6 ensembles, red) and a single-model ensemble (33 IPSL-CM6A-LR members described in Boucher et al., 2020; green). 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.

Due to the expensive computing cost 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 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 for a given expected uncertainty.

In this study, we employ established statistical theory to propose a complementary approach for estimating the required ensemble size for an expected uncertainty. We provide new information about the ensemble uncertainty before the ensemble is generated, enabling those who perform the experiments to decide *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 or to estimate the ensemble size required to reach a given uncertainty of the ensemble mean, 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 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 1850 values. We use 59 ensembles from 53 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). We consider 24 of these ensembles as ‘large ensembles’ (LEs) as they have 10 members or more (for more details about members and ensembles see Text S1 in Supporting Information S1). A multi-model ensemble (hereafter CMIP6-MME) is created using the first member of each 59 ensembles. Monthly means are used for all datasets.

143 Note that we performed a simple quality: i) we computed piControl's global mean
144 surface temperature to verify if the simulated climate is stationary; and ii) we compared the
145 diagnostics (mean, variance and skewness of N3 PR and N3 SST) computed from piControl and
146 the corresponding historical runs to verify if the climate statistics are similar. Following this
147 quality control, simulations of CAS-ESM2-0 and KACE-1-0-G are not used in this study, and the
148 first 650 years of HadGM3-GC31-LL's piControl are also not used (for more details see Text S2
149 and Figure S1 in Supporting Information S1).

150 The epoch 1985-2014 of two observations datasets are used, Global Precipitation
151 Climatology Project Monthly Analysis Product version 2.3 for PR (GPCPv2.3; Adler et al.,
152 2003) and NOAA Optimum Interpolation Sea Surface Temperature version 2 for SST (OISSTv2;
153 Reynolds et al., 2002).

154

155 **Table 1**

156 *List of CMIP6 ensembles, their duration for piControl run and size for historical run*

Model name	Ensemble	PI	HI
ACCESS-CM2	i1p1f1	500	10
ACCESS-ESM1-5	i1p1f1	1000	40
AWI-CM-1-1-MR	i1p1f1	500	5
BCC-CSM2-MR	i1p1f1	600	3
BCC-ESM1	i1p1f1	451	3
CAMS-CSM1-0	i1p1f1	500	2
CanESM5_p1	i1p1f1	1000	25
CanESM5_p2	i1p2f1	1051	40
CanESM5-1	i1p1f1	501	47
CanESM5-CanOE	i1p2f1	501	3
CESM2	i1p1f1	1201	11
CESM2-FV2	i1p1f1	500	3
CESM2-WACCM	i1p1f1	500	3
CESM2-WACCM-FV2	i1p1f1	501	3
CIESM	i1p1f1	500	3
CMCC-CM2-SR5	i1p1f1	500	1
CMCC-ESM2	i1p1f1	500	1
CNRM-CM6-1	i1p1f2	500	29
CNRM-CM6-1-HR	i1p1f2	300	1

Model name	Ensemble	PI	HI
GFDL-ESM4	i1p1f1	500	3
GISS-E2-1-G_p1f1	i1p1f1	851	12
GISS-E2-1-G_p1f2	i1p1f2	650	11
GISS-E2-1-G_p3f1	i1p3f1	601	9
GISS-E2-1-G_p5f1	i1p5f1	501	9
GISS-E2-1-H_p1f1	i1p1f1	801	10
GISS-E2-1-H_p1f2	i1p1f2	451	5
GISS-E2-1-H_p3f1	i1p3f1	451	5
GISS-E2-2-G	i1p3f1	351	5
HadGEM3-GC31-LL	i1p1f3	1350	55
HadGEM3-GC31-MM	i1p1f3	500	4
INM-CM4-8	i1p1f1	531	1
INM-CM5-0	i1p1f1	1201	10
IPSL-CM6A-LR	i1p1f1	2000	33
MCM-UA-1-0	i1p1f1	500	1
MIROC-ES2H	i1p4f2	420	3
MIROC-ES2L	i1p1f2	500	30
MIROC6	i1p1f1	800	50
MPI-ESM-1-2-HAM	i1p1f1	1000	3

CNRM-ESM2-1	i1p1f2	500	10	**MPI-ESM1-2-HR**	i1p1f1	500	10
E3SM-1-0	i1p1f1	500	5	**MPI-ESM1-2-LR**	i1p1f1	1000	50
E3SM-2-0	i1p1f1	500	21	**MRI-ESM2-0**	i1p1f1	701	10
EC-Earth3	i1p1f1	1105	18	NESM3	i1p1f1	500	5
EC-Earth3-AerChem	i1p1f1	500	3	**NorCPM1**	i1p1f1	1500	30
EC-Earth3-CC	i1p1f1	505	10	NorESM2-LM	i1p1f1	501	3
EC-Earth3-Veg	i1p1f1	500	7	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	1
FGOALS-g3	i1p1f1	700	6	**UKESM1-0-LL**	i1p1f2	1880	15
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/.

2.2. Methodology

2.2.1. Diagnostics

The mean ($\bar{x}$; 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). The 30-year epochs are utilized as the reference climate as recommended by the World Meteorological Organization (WMO). This epoch length also roughly corresponds to the overlapping epoch between the satellite era (1980-present) and the CMIP6's historical run (1850-2014). These calculations are done using the CLIVAR ENSO metrics package (Planton et al., 2021), executed via the PCMDI Metrics Package framework (Lee et al., 2023).

2.2.2. *Creating piControl and historical distributions*

For piControl, distributions are created by computing the statistics over non-overlapping epochs. This means that the epoch length influences the number of values in piControl distributions: for a 300-years 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). The statistics are computed independently over a given epoch length every 5 years (e.g., 1850-1879, 1855-1884, 1860-1889, etc.). 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). See Text S3 and Figure S2 in Supporting Information S1 for a detailed demonstration of how the distributions are created.

2.2.3. *Degrees of freedom*

When considering time series, each time step is not fully independent from the others. The number of effectively independent time steps (i.e., number of degrees of freedom) can be estimated using:

$$n^* = \frac{n}{1 + \sum_{i=1}^{L} \rho_i^2} \quad \textbf{(4)}$$

where the autocorrelation function (ρ) is summed over the number of time steps (L) necessary to reach the first two sign changes (e.g., Russon et al., 2014; Atwood et al., 2017).

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:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \textbf{(5)}$$

$$SE_{\sigma^2} = \sigma^2 \sqrt{\frac{2}{n-1}} \quad \textbf{(6)}$$

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

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 and Zwiers, 1999):

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

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 and 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:

$$\Delta = Z \frac{E_\sigma}{\sqrt{N}} \quad \textbf{(9)}$$

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 ($E_{\bar{x}}$), hereafter 'relative uncertainty' ($\Delta_r = 100\,\Delta/E_{\bar{x}}$). However, the relative uncertainty can become minuscule when $E_{\bar{x}} \gg 1$ (e.g., for N3 SST mean; not shown), or gigantic when $E_{\bar{x}} \ll 1$ (e.g., for N3 SSTA skewness; not shown). For simplicity, we use the absolute uncertainty (Δ) in all sections but in section 3.4 in which the relative uncertainty (Δ_r) 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 S4 and Figure S3 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 2 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 here larger

LEs 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 is a total of 188 curves (15 datasets x 9 epoch lengths = 135 for the historical run, and 53 for the piControl run as the ensemble size decreases and fall below the 15 members threshold when the epoch length increases). All 15 datasets align almost perfectly on the theory (dashed black lines) for all three statistical moments computed with N3 PR and N3 SST.

The only notable discrepancy comes from the piControl ensembles of N3 PRA variance computed with the MPI-ESM1-2-LR (yellow-green left-pointing markers in Figure 2c). This is due to the ability of MPI-ESM1-2-LR to simulate extremely rare but extremely large N3 PRA during El Niño events. 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 (not shown).

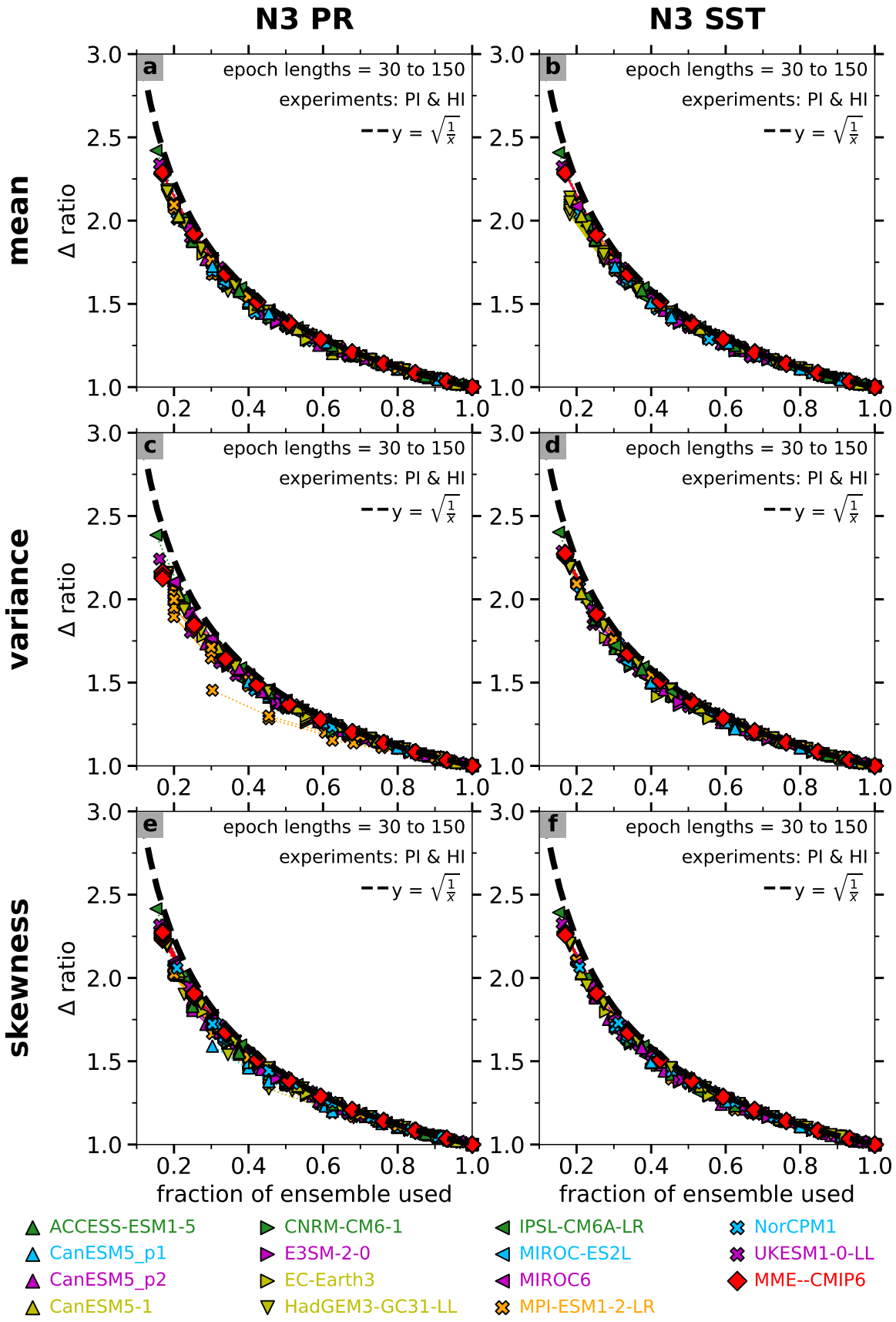

N3 PR
N3 SST
mean
variance
skewness
Δ ratio
a
b
c
d
e
f
epoch lengths = 30 to 150
experiments: PI & HI
--y = √(1/x)
fraction of ensemble used
ACCESS-ESM1-5
CanESM5_p1
CanESM5_p2
CanESM5-1
CNRM-CM6-1
E3SM-2-0
EC-Earth3
HadGEM3-GC31-LL
IPSL-CM6A-LR
MIROC-ES2L
MIROC6
MPI-ESM1-2-LR
NorCPM1
UKESM1-0-LL
MME--CMIP6

Figure 2. Evolution of the uncertainty of the ensemble mean (Δ; equation (9)) as a function of the fraction of the ensemble used. Uncertainty computed for N3 PR (first column) and N3 SST (second column) mean (first row), variance (second row) and skewness (third row). 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

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. For simple diagnostics (like the first three statistical moments), the statistical theory clearly highlights this effect: if one uses equations (5), (6) and (7) with σ and n respectively equal to the standard deviation of the time series and the number of independent time steps (n^*), the three equations indicate a decrease in the error of these statistics 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 3 shows the ratio of the uncertainty of the ensemble mean (Δ) computed with historical ensembles using epoch lengths of 30 to 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 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 24 LEs and the CMIP6-MME, using the maximum number of members of each ensemble (25 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 datasets show an improvement that is broadly consistent with

291 the theory (dashed black lines) for all three statistical moments computed with N3 PR and N3
292 SST.

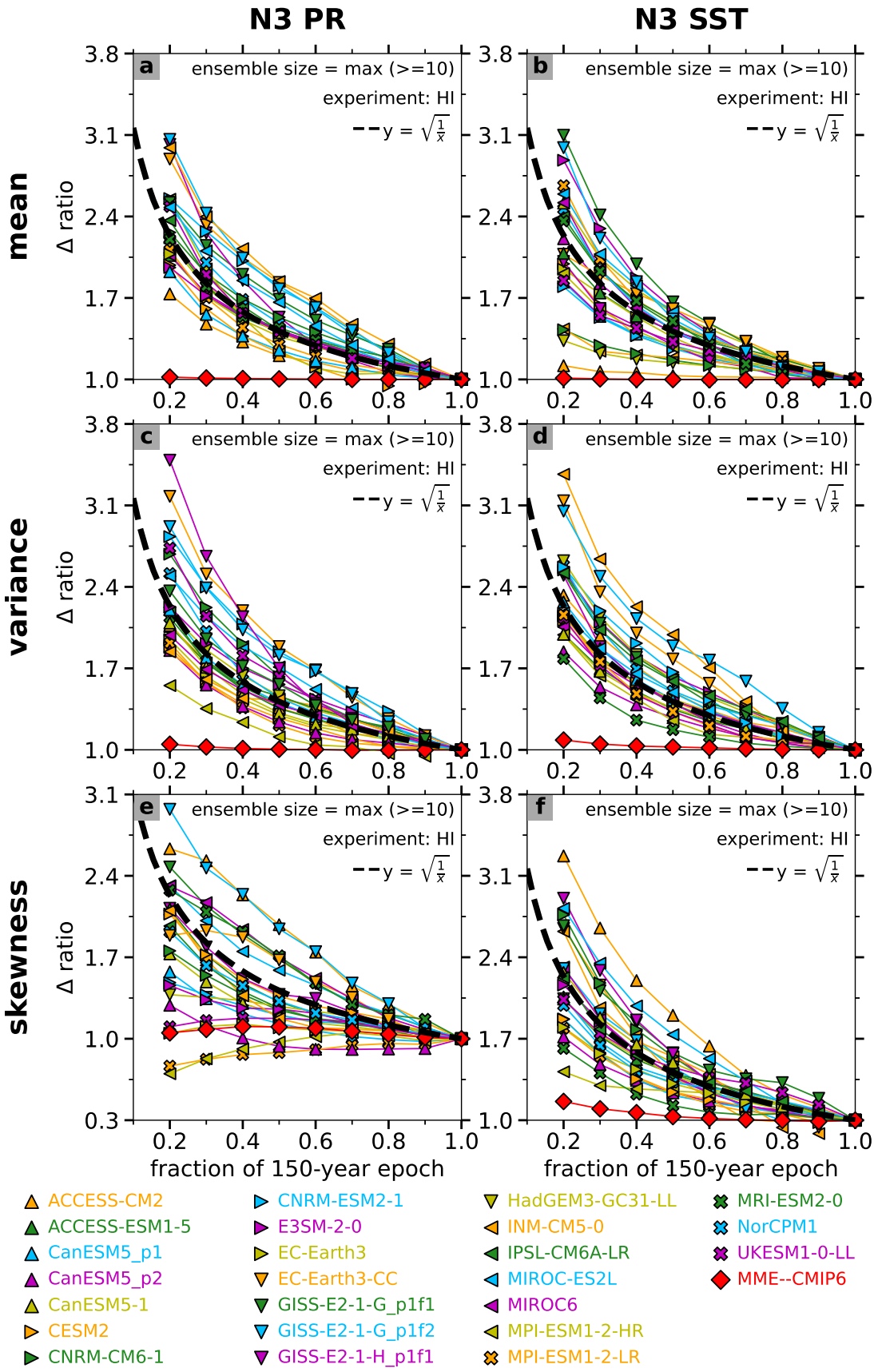
293 However, for N3 SST mean (Figure 3b) and N3 PRA skewness (Figure 3e), several
294 ensembles are clearly departing from the theory. For these ensembles and diagnostics, the intra-
295 ensemble standard deviation (E_σ) is not increasing as fast as expected (or even decreases), with
296 decreasing epoch length. The exact reason is beyond the scope of this paper but three simple
297 reasons may explain this result: i) equations (5), (6) and (7) are valid when the sample is drawn
298 from a normal distribution but N3 PRA and N3 SSTA (to a smaller extent) distributions are
299 skewed (Figure 3j,l); ii) a small sized LE can randomly deviate from the theory (see Text S5 and
300 Figure S4 in Supporting Information S1); and iii) long term trends are not linear (not shown),
301 which is not taken into account by our methodology (time series of each epoch are detrended
302 linearly and independently) and may falsely increase the standard deviations computed with long
303 time series.

304 The behavior of the CMIP6-MME is also notable: varying the epoch length has no
305 influence on the uncertainty. This is due to the fact that increasing the epoch length only
306 attenuates the internal variability within each model, it does not reduce the inter-model
307 differences. So, if one wants to detect a change in a statistical value (e.g., related to climate
308 change) using the CMIP6-MME, increasing the epoch length will not reduce the uncertainty.
309 One may detect a change only if it is large enough between the considered epochs.

310 While the uncertainty should similarly increase with decreasing epoch length in the
311 piControl run, it is not easy to prove it due to the methodology used to create the distributions
312 (see section 2.2.2). Indeed, with the piControl run increasing the epoch length implies a smaller
313 number of samples, reducing our ability to robustly compute the standard deviation of the
314 distribution (E_σ). In addition, for a given epoch length we obtain a single value of the
315 uncertainty, while with the historical run we obtain an uncertainty value for each partially
316 overlapping epoch (e.g., using 30-year epochs we obtain 28 uncertainty values, one for 1850-
317 1879, another for 1855-1884, etc.). Despite these methodological issues, with a long piControl
318 run (~2000 years), the uncertainty of the ensemble mean would follow the theory (not shown).

319 Thus, both ensemble size and epoch length can be used to improve the uncertainty of the
320 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.


N3 PR
N3 SST
mean
variance
skewness
Δ ratio
a
b
c
d
e
f
ensemble size = max (>=10)
experiment: HI
--y = √(1/x)
fraction of 150-year epoch
ACCESS-CM2
ACCESS-ESM1-5
CanESM5_p1
CanESM5_p2
CanESM5-1
CESM2
CNRM-CM6-1
CNRM-ESM2-1
E3SM-2-0
EC-Earth3
EC-Earth3-CC
GISS-E2-1-G_p1f1
GISS-E2-1-G_p1f2
GISS-E2-1-H_p1f1
HadGEM3-GC31-LL
INM-CM5-0
IPSL-CM6A-LR
MIROC-ES2L
MIROC6
MPI-ESM1-2-HR
MPI-ESM1-2-LR
MRI-ESM2-0
NorCPM1
UKESM1-0-LL
MME--CMIP6

325

Figure 3. Evolution of the uncertainty of the ensemble mean (Δ; equation (9)) as a function of the fraction of 150-year used for the computation. Uncertainty computed for N3 PR (first column) and N3 SST (second column) mean (first row), variance (second row) and skewness (third row). 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 24 LEs (using the maximum ensemble size). Note that panel e does not have the same vertical range as the other panels.

3.3. Uncertainty in piControl vs. historical runs

Thompson et al., (2015) proposed that a piControl run 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.

We compare now the uncertainty of the ensemble mean (Δ) computed from the 24 historical LEs and the CMIP6-MME with the corresponding piControl runs (Figure 4), 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-years long, i.e., 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 4b) 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) is much larger than the difference between a member of a given model to another member of the same model (i.e., the mean state modulation by the internal variability).

354 This analysis reveals that four of the six diagnostics (Figure 4a,c,d,e) produce an almost
355 perfect match between historical and piControl runs (correlation > 0.9, slope ~1, intercept ~0).
356 The relationship is not as good in the other two diagnostics (correlation ~0.7, slope ~0.6,
357 intercept > 0; Figure 4b,f,), with better uncertainties in the historical compared to the piControl
358 run when the uncertainty value is large. Overall, the piControl run is a good proxy of the
359 uncertainty of the historical run, meaning that the control simulation can be used when the
360 historical ensemble is small, or to estimate the size of the historical ensemble before computing
361 it. This is useful for modelers because multiple control runs may be performed during the model
362 development or tuning process, well-before historical runs are performed.

363

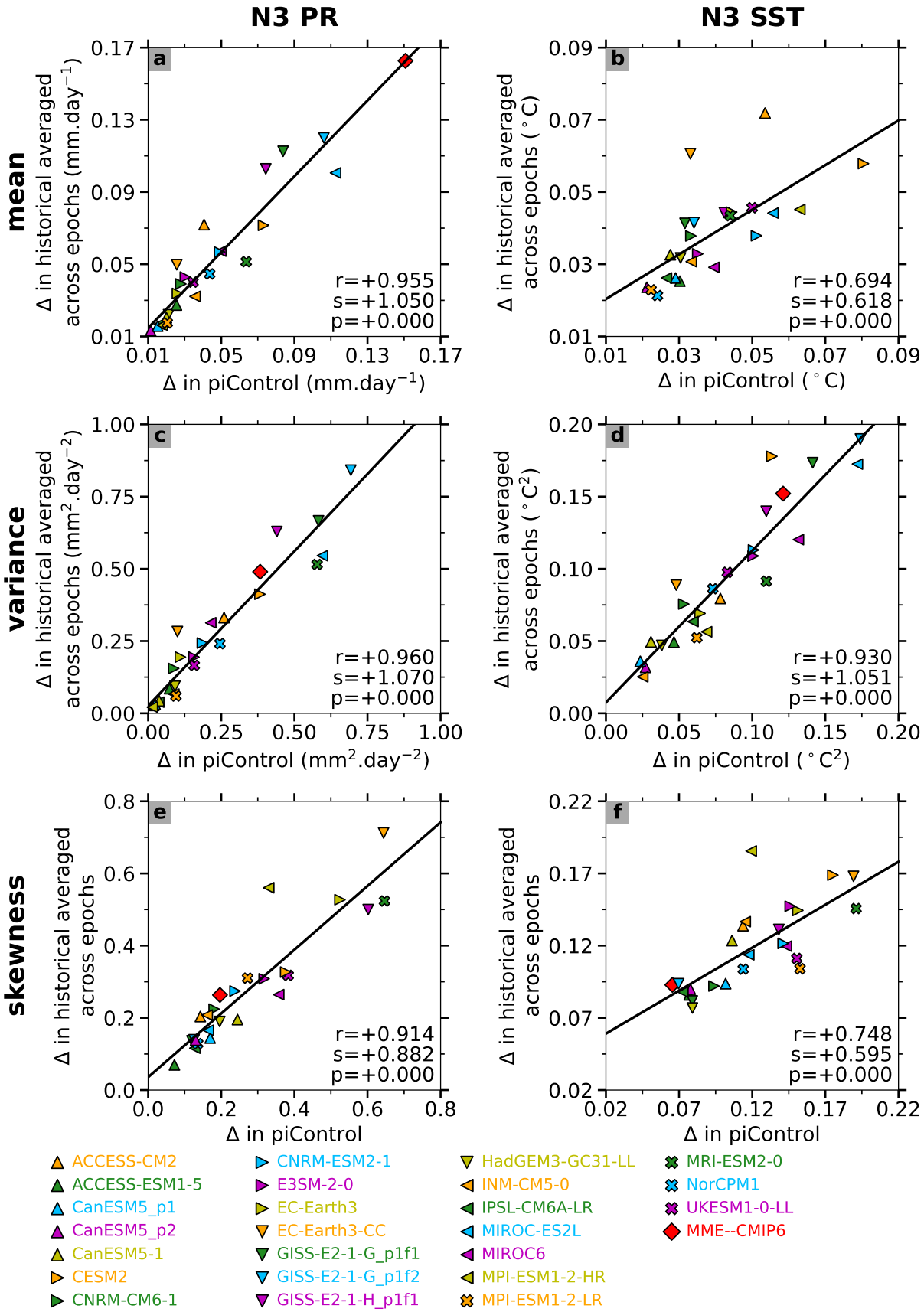

N3 PR
N3 SST
mean
variance
skewness
a
b
c
d
e
f
Δ in historical averaged across epochs (mm.day⁻¹)
Δ in piControl (mm.day⁻¹)
Δ in historical averaged across epochs (°C)
Δ in piControl (°C)
Δ in historical averaged across epochs (mm².day⁻²)
Δ in piControl (mm².day⁻²)
Δ in historical averaged across epochs (°C²)
Δ in piControl (°C²)
Δ in historical averaged across epochs
Δ in piControl
r=+0.955 s=+1.050 p=+0.000
r=+0.694 s=+0.618 p=+0.000
r=+0.960 s=+1.070 p=+0.000
r=+0.930 s=+1.051 p=+0.000
r=+0.914 s=+0.882 p=+0.000
r=+0.748 s=+0.595 p=+0.000
ACCESS-CM2
ACCESS-ESM1-5
CanESM5_p1
CanESM5_p2
CanESM5-1
CESM2
CNRM-CM6-1
CNRM-ESM2-1
E3SM-2-0
EC-Earth3
EC-Earth3-CC
GISS-E2-1-G_p1f1
GISS-E2-1-G_p1f2
GISS-E2-1-H_p1f1
HadGEM3-GC31-LL
INM-CM5-0
IPSL-CM6A-LR
MIROC-ES2L
MIROC6
MPI-ESM1-2-HR
MPI-ESM1-2-LR
MRI-ESM2-0
NorCPM1
UKESM1-0-LL
MME--CMIP6

Figure 4. Uncertainty of the ensemble mean (Δ; equation (9)) computed from historical vs. the piControl runs. Uncertainty computed using 30-year epochs for N3 PR (first column) and N3 SST (second column) mean state (first row), variance (second row) and skewness (third row). Uncertainty computed in the 24 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.

3.4. Estimating the ensemble size

There are many papers in the literature proposing a minimum number of members, often termed the Required Ensemble Size (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 a method that one can apply to estimate the required ensemble size for a particular application, *before* the ensemble is generated. Indeed, the RES can be estimated by rearranging equation (9):

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

Therefore, we 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).

When defining the RES in this section, we 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 (Figure 5b,f; gold), a relative uncertainty (Δ_r) of 5% for N3 PR mean (Figure 5a; gold) and of 20% for N3 PR variance and skewness, as well as N3 SST variance (Figure 5c,d,e; gold). To reach these uncertainties, the IPSL-CM6A-LR ensemble (green 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, implying stronger interdecadal modulation of its variance.

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 dataset (for the 95% confidence interval, $\Delta = P_5|E_{\bar{x}} - obs|$; Figure 5; 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, e.g., 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 ($\Delta = 0.1\ CMIP6's\ IQR$; Figure 5; 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.

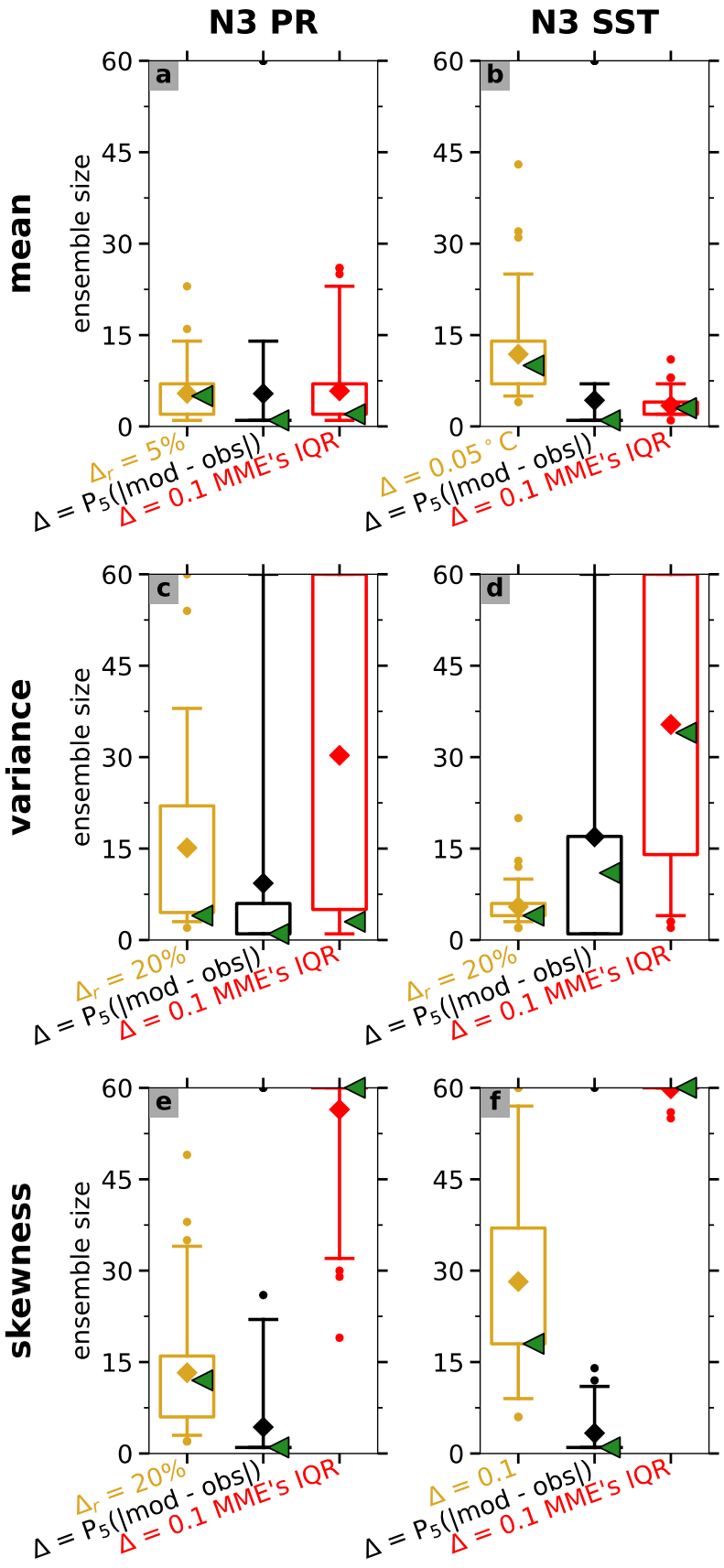

N3 PR
N3 SST
mean
variance
skewness
ensemble size
a
b
c
d
e
f
0
15
30
45
60
Δr = 5%
Δ = 0.05 °C
Δr = 20%
Δ = 0.1
Δ = P5(|mod - obs|)
Δ = 0.1 MME's IQR

Figure 5. Ensemble sizes required to limit the uncertainty to a desired value (equation (10)). 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 ($\Delta = \boldsymbol{P_5}|\boldsymbol{E_{\bar{x}}} - \boldsymbol{obs}|$; black) and to limit the overlap of the confidence interval of each model ($\Delta = \mathbf{0.1}\,\boldsymbol{CMIP6's\ IQR}$; red). RES computed for N3 PR (first column) and N3 SST (second column) mean (first row), variance (second row) and skewness (third row). Green triangles represent the RES for IPSL-CM6A-LR. Boxplots represent the distribution of values computed using all CMIP6 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.

4. Conclusions

We analyzed the first three statistical moments (mean, variance, and skewness) of N3 PR and N3 SST computed on all available CMIP6 piControl and historical ensembles (24 large ensembles and the CMIP6-MME made of the first member from 59 ensembles) 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 according to theory, with the square root of the ensemble size. Thus, if one has an ensemble with an uncertainty Δ, and wishes to reduce it to half Δ, the ensemble size must be quadrupled.
- The epoch length generally has a similar effect on Δ (does not apply to a multi-model ensemble) but there are more inter-model differences, probably linked to the non-normality of the distributions, the relatively small ensemble sizes, and the nonlinearity of climate change trends in simulated historical runs.
- There is a good correspondence between the Δ computed with an historical LE and with the corresponding piControl. This implies that one can use a piControl run to estimate in advance how many historical members must be computed to obtain a given Δ, or to estimate Δ of a small historical ensemble.
- With our methodology one can simply estimate the ensemble size to fit one's purpose, regardless of the ensemble size already computed (if a random sampling is used, as in

458 Milinski et al., 2020, one can estimate the ensemble size only if it is smaller than the one
459 already computed).

460 The methodology that we propose to estimate the ensemble size complements the random
461 sampling performed by Milinski et al., (2020) and J. Lee et al., (2021), but at a much smaller
462 computation cost (no random selections). As we provide the mathematical formulae to compute
463 the uncertainty of an ensemble mean (equation (9)) or the ensemble size required to reach a
464 given uncertainty of the ensemble mean (equation (10)), our results have numerous advantages.
465 Our equations can be used by any model user to fit their own purpose. One can also extrapolate
466 their results: using a computation done with a given ensemble size and a given epoch length one
467 can estimate the uncertainty for other ensemble sizes or epoch lengths. And finally, we used
468 simple statistics to illustrate how statistical theory can be applied to climate science, but
469 equations (9) and (10) can be used for any diagnostic using any variable if the distributions are
470 approximately normal.

471

472 **Acknowledgments**

473 We acknowledge the World Climate Research Programme's Working Group on Coupled
474 Modelling, which is responsible for organizing CMIP. We thank all the international climate
475 modeling groups for their tireless development efforts, and for generously producing and
476 publishing these coordinated, standardized, and quality-controlled simulations. We thank the
477 Earth System Grid Federation (ESGF) for archiving the simulations and improving access, and
478 are grateful to the multiple funding agencies who support CMIP and ESGF. The U.S.
479 Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides
480 coordinating support for CMIP, and led development of software infrastructure in partnership
481 with the Global Organization for Earth System Science Portals. Author Planton held a National
482 Research Council Research Associateship at NOAA/ PMEL when he started this work. He is
483 now supported by the the Australian Government's National Environmental Science Program
484 (NESP2) Climate Systems Hub. Author McGregor was supported by NESP2 Climate Systems
485 Hub and the Australian Research Council (grant numbers FT160100162 and DP200102329).
486 Work of LLNL-affiliated authors was performed under the auspices of the U.S. Department of
487 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).

Open Research

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/. Datasets 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., 2023; https://github.com/PCMDI/pcmdi_metrics). The output and processing scripts used for the paper are available at https://github.com/yyplanton/estimating_uncertainties_enso.

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

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

Adler, R.F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., & Arkin, P. (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

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. *FAO Fisheries and Aquaculture Technical Paper No. 660*, Rome, FAO. https://doi.org/10.4060/ca8348en

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D'Andrea, F., Davini, P., de Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J.-L., Dupont, E., Éthé, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols, M.-A., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J.-Y., Guenet, B., Guez, L. E., Guilyardi, E., Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas, N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J.-B., Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P., Planton, Y. Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P., Sima, A., Swingedouw, D., Thiéblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard, J., Viovy, N., & Vuichard, N. (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

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

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: Princeton University Press. https://doi.org/10.1515/9781400883868

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson I. R., & Ting M. (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., Doutriaux, C., Stockhause, M., & Gleckler, P. J. (2018). Toward standardized data sets for climate model experimentation. *Eos*, *99*. 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

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

Lee, J., Gleckler, P. J., Ahn, M.-S., Ordonez, A., Ullrich, P. A., Sperber, K. R., Taylor, K. E., Planton, Y. Y., Guilyardi, E., Durack, P., Bonfils, C., Zelinka, M. D., Chao, L.-W., Dong, B., Doutriaux, C., Zhang, C., Vo, T., Boutte, J., Wehner, M. F., Pendergrass, A. G., Kim, D., Xue, Z., Wittenberg, A. T., & Krasting, J. (2023). Objective Evaluation of Earth

System Models: PCMDI Metrics Package (PMP) version 3, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2023-2720

Lee, J., Planton, Y. Y., Gleckler, P. J., Sperber, K. R., Guilyardi, E., Wittenberg, A. T., McPhaden, M. J., & Pallotta, G. (2021). 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., Engelbrecht, F., Fischer, E., Fyfe, J. C., Jones, C., Maycock, A., Mutemi, J., Ndiaye, O., Panickal, S.,& Zhou, T. (2021). 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, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, B. Zhou (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, Cambridge, United Kingdom and New York, NY, USA. 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., 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., Stouffer, R. J., & Taylor, K. E. (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

Planton, Y. Y., Guilyardi, E., Wittenberg, A. T., Lee., J., Gleckler, P. J., Bayr, T., McGregor, S., McPhaden, M. J., Power, S., Roehrig, R., Vialard, J., & Voldoire, A. (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

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

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

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

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

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., Samuelsen, A., Langehaug, H., Svendsen, L., Chiu, P.-G., Passos, L., Bentsen, M., Guo, C., Gupta, A., Tjiputra, J., Kirkevåg, A., Olivié, D., Seland, Ø., Solsvik Vågane, J., Fan, Y., & Eldevik, T. (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., Ackerman, A. S., Aleinov, I, Bauer, M., Bleck, R., Canuto, V., Cesana, G., Cheng, Y., Clune, T. L., Cook, B. I., Cruz, C. A., Del Genio, A. D., Elsaesser, G. S., Faluvegi, G., Kiang, N. Y., Kim, D., Lacis, A. A., Leboissetier, A., LeGrande, A. N., Lo, K. K., Marshall, J., Matthews, E. E., McDermid, S., Mezuman, K, Miller, R. L., Murray, L. T., Oinas, V., Orbe, C., García-Pando, C. P., Perlwitz, J. P., Puma, M. J., Rind, D., Romanou, A., Shindell, D. T., Sun, S., Tausnev, N., Tsigaridis, K.,Tselioudis, G., Weng, E., Wu, J., Yao, M.-S. (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., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., von Salzen, K., Yang, D., & Winter, B. (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