Journal of Advances in Modeling Earth Systems

Supporting Information for

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

Contents of this file

Introduction

In this supporting information, we describe details about CMIP6 ensembles (Text S1), and the quality control that we performed on CMIP6 data (Text S2 and Figure S1). Then we compare the methodology used in the paper to compute the uncertainty of the ensemble mean (based on the statistical theory) to the random sampling often used in the literature (e.g., Milinski et al., 2020; Text S3 and Figure S2). One model (MPI-ESM1-2-LR) simulates rare extremely large precipitation anomalies making it deviate from the theoretical relationship between uncertainty and ensemble size (blue cross markers of Figure 3d of the main manuscript). Figure S3 shows that by removing these extremes this model gets closer to the theory. In the main manuscript we argued that several large ensembles (LEs) are still small and can randomly deviate from the theoretical relationship between uncertainty and epoch length. This is demonstrated using subsamples of one of the largest

LE available (Text S4 and Figure S4). Evaluating the influence of the epoch length on uncertainty using piControl runs is delicate as the number of samples decreases with increasing epoch length. Nevertheless, the uncertainty decreases with the epoch length, broadly following the theory (Text S5 and Figure S5). Finally, we compare the ensemble size required to limit the uncertainty to a desired value computed with our method based on the statistical theory and with random sampling (Text S6 and Figure S6).

Text S1. CMIP6 ensembles

The CMIP6 historical run is composed of one or more members branching from the piControl run. For most climate models, a historical ensemble is generated by starting members from initial conditions sampled from the piControl run at regular intervals (e.g., IPSL-CM6A-LR; Boucher et al., 2020). For some climate models (e.g., NorCMP1; Bethke et al., 2021), the historical ensemble is initialized from the piControl by adding small random noise at the beginning, integrated for 100 years to allow the spread to grow, and after this integration period the historical run is started.

CMIP6 members are named using 4 indices ("r" for realization, "i" for initialization procedure, "p" for physics, and "f" for forcing), each followed by a number (e.g., r1i1p1f1). In this paper, historical ensembles are created using varying realization ("r" value is changing) but constant initialization procedure, physics and forcing ("i", "p" and "f" are the same). This means that members r1i1p1f1 and r2i1p1f1 will form an ensemble but if another member r1i1p1f2 is available, it will be placed in another ensemble.

Models CanESM5, GISS-E2-1-G and GISS-E2-1-H have multiple ensembles as multiple physics and forcings are tested on them. In CanESM5 p1 a conservative remapping is used when the wind stress field is passed from the atmospheric model to the oceanic model, while in p2 bilinear regridding is used (Swart et al., 2019). Concerning GISS models, GISS-E2-1-G and GISS-E2-1-H are considered different models as the ocean model is different: GISS Ocean for "G" and HYCOM for "H". Both models have several ensembles: those named "p1" a noninteractive ozone and aerosol physics is used while in "p3" and "p5" a chemistry model is used (respectively OMA and MATRIX models); "f2" is a correction of "f1", mainly regarding volcanic aerosol (Kelley et al., 2020).

Text S2. Quality control

Before analyzing the CMIP6 data for this paper we performed a simple quality control. First, we computed the Global Mean Surface Temperature (GMST) of the piControl run to verify if the Earth's climate is stationary. Figure S1a shows the evolution of the GMST (average of the first 30-years removed) computed from the piControl run of 61 CMIP6 ensembles. It clearly shows that the KACE-1-0-G's GMST (red curve) is steadily increasing (~0.4°C/century). One can also note that HadGM3-GC31-LL's GMST (green curve) increases between years 450 and 650 (~0.3°C/century) and then stabilizes around a temperature ~0.6°C warmer than at the beginning of the simulation. Then we compared the diagnostic values (mean, variance and skewness of N3 PR and N3 SST) computed from

the piControl run and the corresponding historical runs to verify if both types of runs have relatively similar climate statistics. Figure S1b-g shows the six diagnostics (averaged piControl value removed) computed from piControl and historical runs of 61 CMIP6 ensembles. The distributions are quite different, but they always overlap, but for the N3 PRA variance of CAS-ESM2-0 (Figure S1d).

For these reasons, 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.

Text S3. Uncertainties of the ensemble mean: theory vs. random sampling

The uncertainty of the intra-ensemble mean (Δ) computed using equation (9) should be equivalent to that computed using random sampling. For the latter, a nonparametric Monte Carlo method is used. For each ensemble distribution of size *N*, *N* members are randomly selected, with replacement, and the mean of this resampled ensemble is calculated. This operation is repeated 1,000,000 times, yielding a distribution of absolute deviations from the full ensemble mean. The 2.5^{th} and 97.5^{th} percentiles of this distribution are used to define the 95% confidence interval on the intra-ensemble mean. Figure S2 compares the uncertainty of the intra-ensemble mean (Δ) computed with both methods, using all available ensembles with at least 10 members (piControl run from all ensembles, the 24 historical LEs and the CMIP6-MME), and using epoch lengths ranging from 30 to 150 years (15-year intervals). The two methods yield almost identical results: correlation and slope equal to 1, p-value < 0.001. Some small differences appear for the large uncertainties of the ensemble mean of N3 PR skewness (Figure S2e), with theoretical values (vertical axis) slightly larger than that computed with random sampling (horizontal axis).

Text S4. Ensemble size and influence of the epoch length on the uncertainty

We established in the main manuscript that, for the mean, variance and skewness, the uncertainty of the intra-ensemble mean (Δ) should decrease with the square root of the epoch length. However, we showed that some models are not perfectly following this relationship. Part of this mismatch is caused by the relatively small size of several ensembles. This can be demonstrated by computing synthetic ensembles of reduced size (using combinations) from one of the largest ensembles (ACCESS-ESM1-5, 40 members), to provide a range of values that this ensemble could take if fewer members had been computed (Figure S4). The wide range of values taken by the synthetic ensembles of ACCESS-ESM1-5 indicate that an ensemble can randomly deviate from the theory if the ensemble is too small.

Text S5. piControl and influence of the epoch length on the uncertainty

As discussed in the main manuscript, for piControl runs non-overlapping epochs are used to compute the statistics (to obtain independent measurements) implying that the number of samples reduces as the epoch length increases. For example, with the 2000-year piControl simulation of IPSL-CM6A-LR, there are 66 non-overlapping epochs of 30

years, but only 16 of 120 years. This implies that to estimate the influence of the epoch length on the uncertainty in piControl runs, we need to use combinations of k distinct samples, k being the smallest sample size. In the case of IPSL-CM6A-LR, $k = 16$ if we analyze the evolution of the uncertainty of the ensemble mean to that computed with 120-year epochs. This means that for piControl runs, we are only using relatively small sample sizes, implying that the results could randomly deviate from the theory (Text S4). Nevertheless, the uncertainty of the ensemble mean computed with the longest piControl runs (Figure S5) is decreasing with increasing epoch length, broadly consistent with the theory.

Text S6. Required ensemble size: theory vs. random sampling

To estimate the required ensemble size (RES), one can use equation (10), or generate random samples. For the latter, a nonparametric Monte Carlo method is used, as in J. Lee et al. (2021). For each ensemble distribution, k members are randomly selected, with replacement, and the mean of this resampled ensemble is calculated and compared to the mean computed with the full ensemble. This operation is repeated 1,000,000 times, yielding a distribution of absolute deviations from the full ensemble mean. The 2.5^{th} and 97.5^{th} percentiles of this distribution are used to define the 95% confidence interval on the intra-ensemble mean. By incrementing k from 1 to the full ensemble size (N), one can find how many members are required to reach a given uncertainty. Figure S6, comparing the two methodologies, shows that they are perfectly equivalent. Note that the RES computed by random sampling is limited by the available ensemble size (N): if the desired uncertainty cannot be reached with N members, the RES cannot be defined. This is not the case if equation (10) is used.

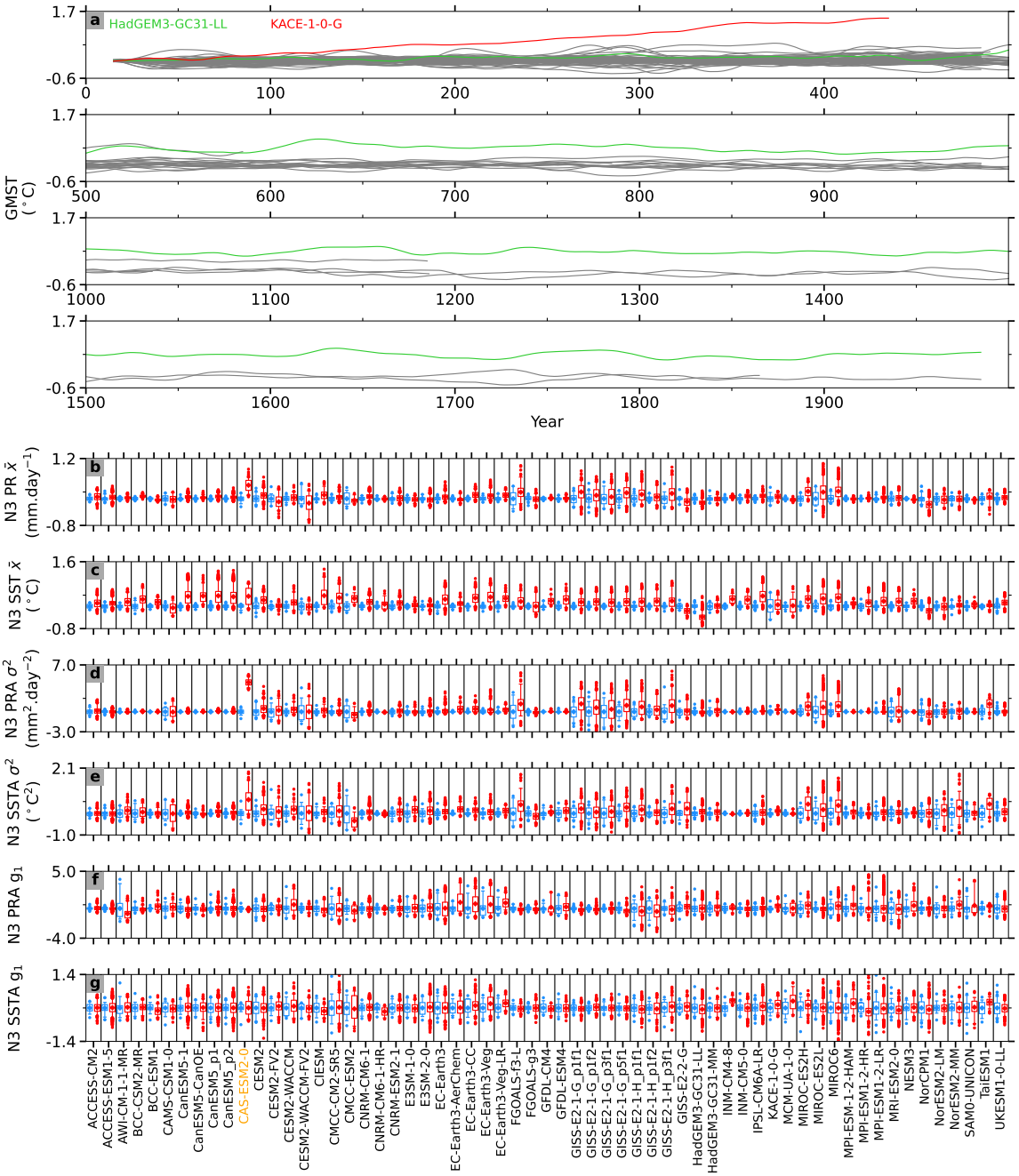

Figure S1. Quality control of CMIP6 data. (a) Time series of Global Mean Surface Temperature (GMST) computed from piControl runs of 61 CMIP6 ensembles (GMST averaged over the first 30-years removed). Values of (b) N3 PR mean ($\bar{\boldsymbol{x}}$), (c) N3 SST $\bar{\boldsymbol{x}}$ (both computed using equation (1)), (d) N3 PRA variance ($\boldsymbol{\sigma^2}$), (e) N3 SSTA $\boldsymbol{\sigma^2}$ (both computed using equation (2)) and (f) N3 PRA skewness ($\boldsymbol{g_1}$), (g) N3 SSTA $\boldsymbol{g_1}$ (both computed using equation (3)) computed from 30-year epochs of piControl runs (blue) and historical runs (red; all members and epochs) of 61 CMIP6 ensembles (piControl average removed). b-g) 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.

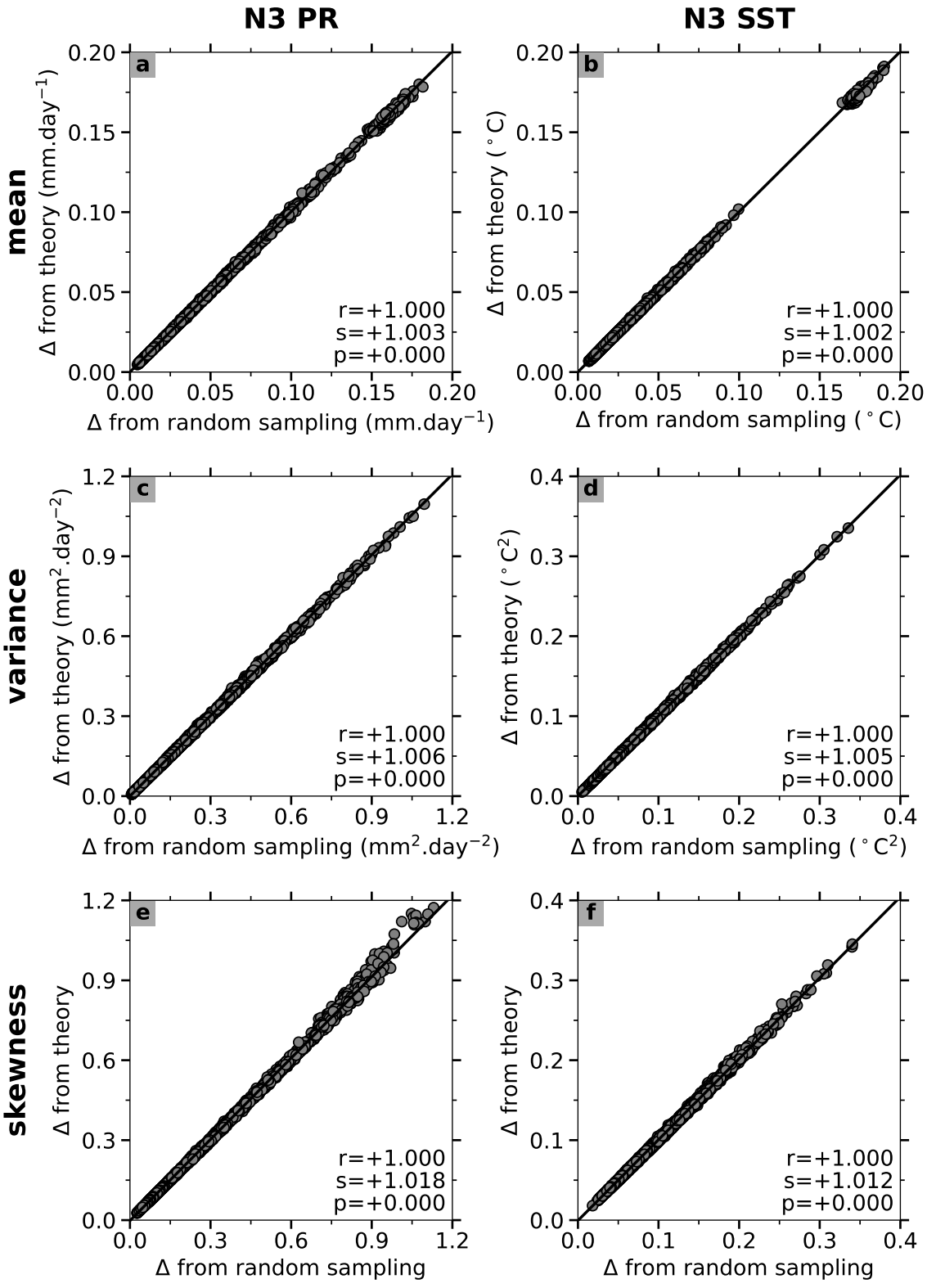

Figure S2. Comparison of the uncertainty of the intra-ensemble mean (Δ) computed using equation (9) and using random sampling. The uncertainty is computed for N3 PR (first and

third columns) and N3 SST (second and fourth columns) mean (first row), variance (second row) and skewness (third row) from piControl runs (59 ensembles and CMIP6-MME) and historical runs (26 LEs and CMIP6-MME), epoch lengths ranging from 30 to 150 years (every 15 years), all epochs. The black line represents the linear regression slope, and the corresponding correlation (r), slope (s) and p-value (p) are indicated at the bottom of the panel.

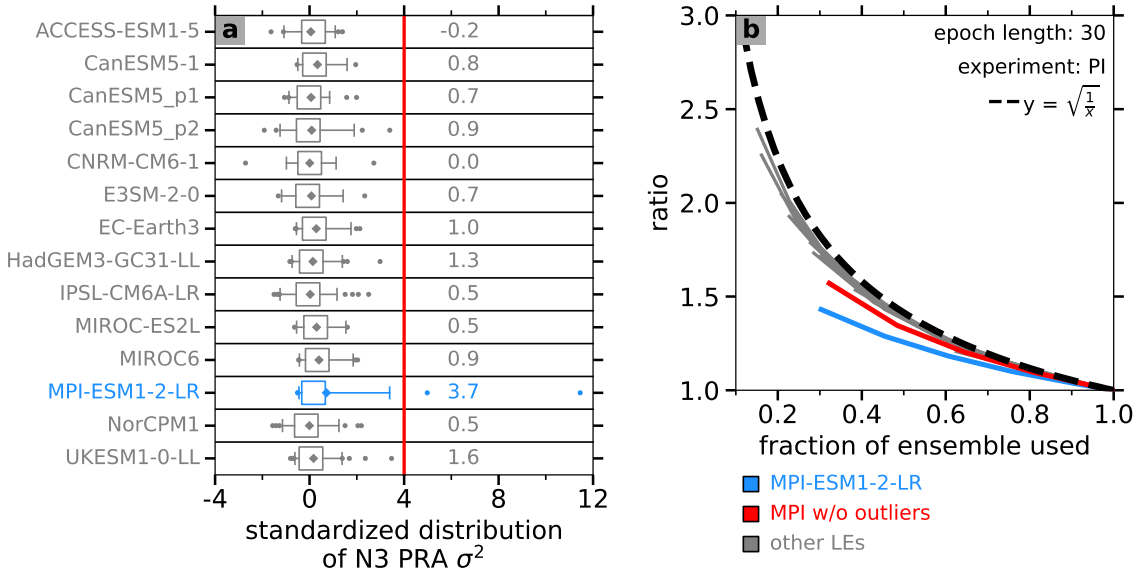

Figure S3. Impacts of outliers on the evolution of the uncertainty of the ensemble mean (Δ; equation (9)). (a) Standardized distributions (median removed and divided by the interquartile range) of N3 PRA variance and (b) dependence of the uncertainty of the ensemble mean on the fraction of the ensemble used. 14 piControl runs at least 450 years long (i.e., 15 non-overlapping epochs of 30 years) are used. (a) 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. The skewness ($\boldsymbol{g_1}$; equation (3)) of each distribution is indicated to the right of each distribution. (b) The dashed black, solid blue, red and grey lines represent respectively the theoretical improvement of the uncertainty with the square root of the fraction of the ensemble used, the MPI-ESM1-2-LR using all available values, the MPI-ESM1-2-LR without the two outliers (value > 4; red line in panel (a)), and all the other datasets.

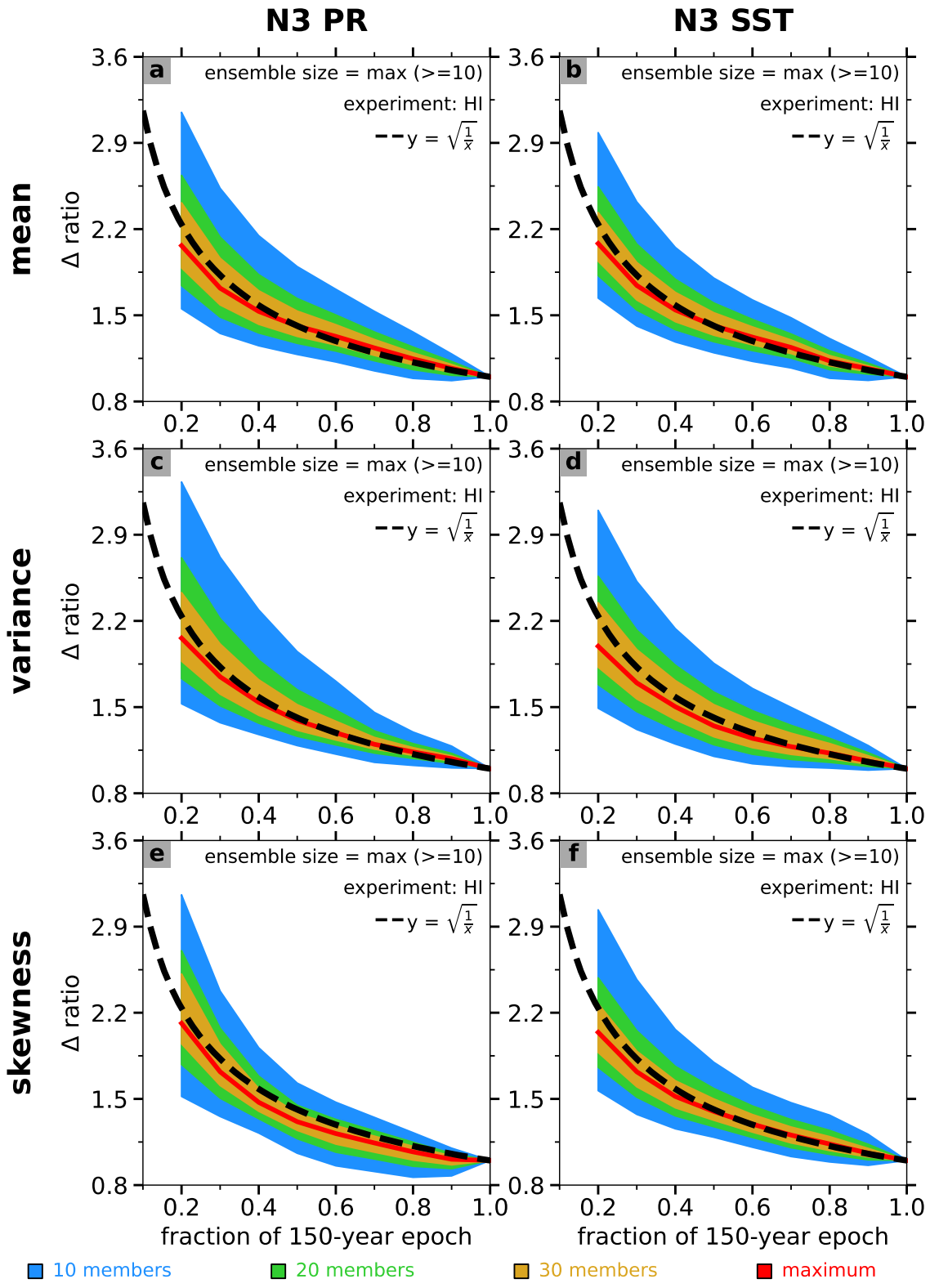

Figure S4. Ensemble size influencing the relationship between the uncertainty of the ensemble mean (Δ; equation (9)) and the epoch length. Uncertainty computed for N3 PR

(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 ACCESS-ESM1-5 using the full ensemble (red) and combinations of members to estimate the 95% confidence interval of the evolution of Δ with the epoch length if the ensemble had only 10 (blue), 20 (green) or 30 (gold) members.

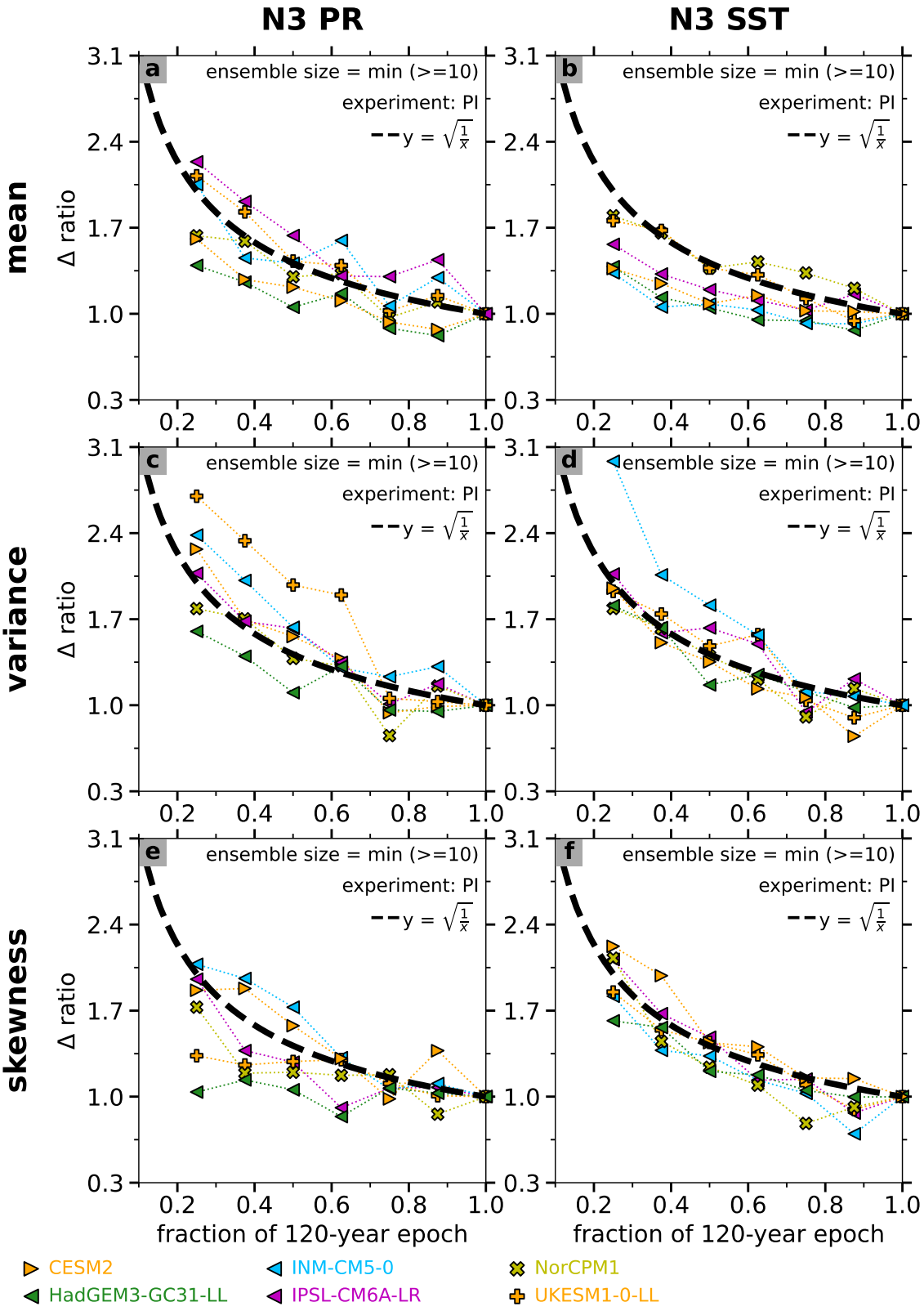

Figure S5. Dependence of the uncertainty of the ensemble mean (Δ; equation (9)) on the fraction of 120-year used for the computation. Uncertainty computed for N3 PR (first

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 120-year used. Uncertainty computed with the longest piControl run from 6 ensembles, using the combinations of the minimum number of samples: CESM2 (10), HadGEM3-GC31-LL (10), INM-CM5-0 (10), IPSL-CM6A-LR (16), NorCPM1 (12), UKESM1-0-LL (15).

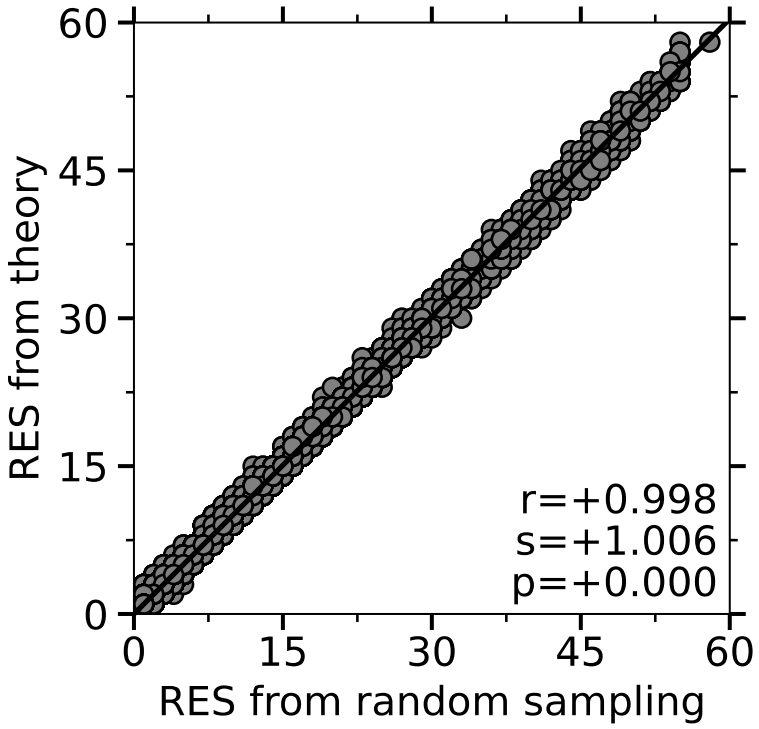

Figure S6. Comparison of the required ensemble size (RES) computed using equation (10) and using random sampling (as in Milinski et al., 2020; J. Lee et al., 2021). The RES is computed for the mean, variance and skewness of N3 PR and N3 SST from piControl runs (59 ensembles and CMIP6-MME) and historical runs (24 LEs and CMIP6-MME), epoch lengths ranging from 30 to 150 years (every 15 years), all epochs, using relative uncertainties (Δ_r) ranging from 5% to 100% (every 5%) for all diagnostics but N3 SST mean, for which relative uncertainties are ranging from 0.1% to 1% (every 0.1%). The black line represents the linear regression slope, and the corresponding correlation (r), slope (s) and p-value (p) are indicated at the bottom of the panel.