Supplementary Material for:

The forced response of the El Niño–Southern Oscillation-Indian monsoon teleconnection in ensembles of Earth System Models

Tamás Bódai$^{1,2,3,4}$, Gábor Drótos$^{5,6,7}$, Mátyás Herein$^{6,8,*}$, Frank Lunkeit$^8$, Valerio Lucarini$^{3,4,8,9}$

$^1$Pusan National University, Busan, Republic of Korea
$^2$Center for Climate Physics, Institute for Basic Science, Busan, Republic of Korea
$^3$Department of Mathematics and Statistics, University of Reading, Reading, UK
$^4$Centre for the Mathematics of the Planet Earth, University of Reading, Reading, UK
$^5$Instituto de Física Interdisciplinar y Sistemas Complejos, CSIC-UIB, Palma de Mallorca, Spain
$^6$MTA–ELTE Theoretical Physics Research Group, and Institute for Theoretical Physics, Eötvös University, Budapest, Hungary
$^7$Max-Planck-Institut für Meteorologie, Hamburg, Germany
$^8$CEN, Meteorological Institute, University of Hamburg, Hamburg, Germany
$^9$Walker Institute for Climate System Research, University of Reading, Reading, UK

*hereinm@gmail.com
Part I. Initialization and convergence

The CESM-LE has initial conditions obtained as minor perturbations of a single trajectory (called “micro initialization”: http://www.cesm.ucar.edu/projects/community-projects/MMLEA/), i.e., very localized in the phase space. It is therefore “obligatory” to discard the beginning of the simulation, until the trajectories sample the natural probability distribution correctly. A safe estimate for this convergence is 40 years, which originates from the discussion of the relevant time scales by Kay et al. (2015) and from additional CESM results (Kim et al., 2017).

As for the MPI-ESM, we suppose, on the one hand, that the relevant time scales are similar in different climate models of similar complexity in similar setups, and, on the other hand, preliminary results indeed suggest that the duration of a primary convergence should be considerably shorter than 40 years. Furthermore, the initialization scheme is very favorable in the MPI-GE (called “macro initialization”).

In particular, the initialization of each of the 63 members in the MPI-HE and of the 43 members in the MPI-1pctE is done by picking a particular time instant from a 1000-year-long pre-industrial control run, in which external conditions are kept at the 1850 level eternally. The different members thus sample the complete attractor (disregarding the variables of the deep ocean, cf. the last paragraph of Section 2 of the main text) corresponding to 1850, but they might not be perfectly uncorrelated from each other, since the average time between their initializations is around 20 years, which is less than the 40-year “safety limit” for the convergence and for the corresponding memory loss. As the total length of the preindustrial run, 1000 years, is much longer than 40 years, the preindustrial attractor is nevertheless sampled correctly by the initial conditions, and their potential correlation can be interpreted as a smaller effective size of the ensembles at the beginning of the simulations (up to e.g. 1890).

(We note that the intended length of the preindustrial control run is 2000 years, with 37 additional ensemble members were initialized in its first half. However, this period proved to be concerned by spin-up effects e.g. in the ocean (Maher et al., 2019). Accessible data for 20 further members is missing from the MPI-1pctE for unknown reason.)

While the above-mentioned potential correlation between different members is one reason to skip this initial period for the computations, a more important reason is that the stationary attractor of the preindustrial control run, obtained by assuming conditions of 1850 eternally, is not identical to the snapshot attractor of 1850 that is determined by the historical forcing scenario (Lamarque et al., 2010) before 1850. Although this difference is supposed to be small, it follows from its presence that also afterwards, in the initial years of the simulation, the members of the MPI-HE are not yet distributed according to the natural distribution of the historical forcing scenario. This is so only after convergence takes place to the latter distribution (i.e., the 1850 initial conditions are forgotten). Our “safe estimate” is 1890 for the time when the convergence is ready. We believe that our argument is a strong theoretical one for discarding the period preceding 1890 (or possibly some earlier date resulting from some thorough analysis) from the MPI-HE. For consistency and comparability, it seems to be reasonable to discard this period from the MPI-1pctE, too. However, as for the MPI-RCP8.5E and the MPI-RCP2.6E, the initial conditions of their members are the endpoints of the trajectories of the MPI-HE, which have already converged by 1890, i.e., earlier than the beginning of the MPI-RCP8.5E and the MPI-RCP2.6E in 2006. Therefore, discarding of initial years is not needed in the latter ensembles.
Part II. The Southern Oscillation Index in a changing climate

The Southern Oscillation Index (SOI) is one of the most important climate indices; it is used to detect changes in ENSO both for the past and in predictions (Power and Kociuba, 2011). There are different definitions for the SOI, but all of them agree in using temporal averages. For simplicity, let us take the station-based definition by the Bureau of Meteorology of the Australian Government (BOM), which is also called the Troup SOI (Troup, 1965):

$$SOI = 10 \frac{p_{\text{diff}}(t) - \bar{p}_{\text{diff}}(t)}{\sqrt{\bar{p}_{\text{diff}}(t)^2 - p_{\text{diff}}(t)^2}}, \quad (S1)$$

Here $p_{\text{diff}}$ is the difference between the mean sea level pressures at Tahiti and Darwin for a particular month (in our paper, we allow for seasonal means as well). The overbar denotes long-term average over some fixed interval of time (e.g. between 1920 and 1950). What is called a La Niña (El Niño) phase corresponds to a positive (negative) value of the SOI if its magnitude exceeds 7 according to BOM.

The problem with (S1) is two-fold. First, the time averages are constants, so that $p_{\text{diff}}$, the only time-dependent term, includes climatic trends instead of characterizing solely anomalies with respect to the instantaneous climatic mean (which is changing in time itself). This problem is illustrated well by considering different climatologies, i.e., taking the temporal averages over different time intervals: it turns out that the values of SOI can be dramatically misleading. Supplementary Fig. S1 shows that we obtain several years when we can identify even both La Niña or El Niño phase depending on the applied climatology. See also Supplementary Discussion I of Herein et al. (2017).

Although there exist sophisticated methods for removing trends from time series, they can resolve the problem only approximately without an a priori knowledge of what should be identified as a trend (i.e., how the real expectation value of a given quantity evolves in time). Furthermore, the experience of Herein et al. (2016) and Herein et al. (2017) indicates that time averages of relevant quantities taken over single time series are influenced by internal variability too much to be able to represent expectation values faithfully. Note that both problems are present for any traditional definition of SOI (or that of any climate index), including those that standardize the sea-level pressures first and take the difference afterwards (e.g. Trenberth, 1976; 1984).

All conceptual problems are resolved, however, by a new, ensemble-wise SOI (which we denote by SOI$_E$):

$$SOI_E = 10 \frac{p_{\text{diff}}(t) - \langle p_{\text{diff}}(t) \rangle}{\sqrt{\langle p_{\text{diff}}(t)^2 \rangle - p_{\text{diff}}(t)^2}}, \quad (S2)$$

where $\langle \ldots \rangle$ denotes averaging with respect to the ensemble in the given time instant $t$ (but only after convergence took place). Evaluating the averages as such ensures the incorporation of the correct properties of the underlying probability distribution. In particular, SOI$_E$ gives the deviation of $p_{\text{diff}}$ of one given realization (note that this is the modeling equivalent of an instrumental record) from the expectation value of $p_{\text{diff}}$, normalized by the ensemble-wise standard deviation. This is so in any year, as a consequence of which a natural detrending is
provided: the climatic mean of SOI\textsubscript{E} is always zero, and the climatic standard deviation of it is always unity times 10.

Note that due to the perpetual zero mean and constant standard deviation, signatures of climate change may be observed only in higher moments of ensemble-wise indices or anomalies, like (S2), so that shifts towards a particular phase or sign cannot exist in the sense of averages. On the contrary, climate change (a response to external forcing) is obviously detectable in ensemble means of non-detrended quantities, see e.g. Supplementary Fig. S2.

Supplementary Fig. S1. The traditional Troup SOI (S1) for the month of November, in the first realization of CESM-LE, as a function of time. Panel (a) shows the SOI calculated with a standard climatology (1920-1950), panel (b) shows the same with a different climatology (2070-2100). For the climatology the model data have been used.

Supplementary Fig. S2. The November sea level pressure difference (p\textsubscript{diff}) between Tahiti and Darwin versus time, in the first realization of CESM-LE (blue), and after averaging over the ensemble instantaneously (red). Grey color indicates all further members of the 35-member ensemble of CESM-LE. The ensemble average shows an enhanced after the year 2050.
Part III. Accommodating correlations in the Mann-Kendall test

The original Mann-Kendall test (Mann, 1945) assumes no correlations in the time series. A modified Mann-Kendall test was developed by Hamed and Rao (1998) that relaxes this assumption. However, the application of the modified test results in p-values of the same order of magnitude as that of the original test for all ensembles, which does not alter the significance of the test result in any of the cases. In what follows, we shall concentrate on the full length of the MPI-HE and MPI-RCP8.5E stitched together and the option of representing the ENSO by the SOI, since for this we obtained a rather counterintuitive result, calling for a close examination of the corresponding p-value, which is $p = 4 \times 10^{-4}$. (We have carried out the below computation for Niño 3 as well, ending up with the same conclusion.)

Obtaining very similar p-values by the original and the modified tests is clearly to do with very weak correlations in the time series, if any. This is indicated by a straightforward calculation of the temporal autocorrelation function displayed in Supplementary Fig. S3 (a). For this we employed the Matlab function ‘autocorr’. Note, however, that the usual autocorrelation function evaluated by an integral over time is well-defined only in the case of stationary processes. In the presence of a trend the estimated correlations are, in principle, not meaningful. Fortunately, the approximate shape of the investigated distribution (a Gaussian) and its standard deviation ($1/\sqrt{N-3}$, where $N$ is the ensemble/sample size) are constant (Fisher, 1936), so that a detrending of the mean of the distribution would transform the time series to that of a stationary process.

However, it is not possible to correctly detrend the data, because the signal that we need to subtract is unknown. In fact, this is the signal of central interest, and all that we attempt is to decide whether it is very likely not stationary, i.e., not constant. Nevertheless, when differences in the subsequent data points in the noisy signal (where noise is due to the finite size of the ensemble in our case) are much bigger than the corresponding differences in the true signal, then differencing (i.e., numerically differentiating) naturally results in a well-detrended signal. Applying this assumption is prompted to be correct by the fact that the sample standard deviation of the $z$ signal (calculated over time) is measured to be 0.133, while the true value for a stationary $z$, calculated as $1/\sqrt{N-3}$, $N = 63$, would be 0.129, which is very close to the previous value.

It can be shown easily that the differencing of an uncorrelated stationary signal leads to a $-1/2$ lag-1 autocorrelation. A value of about $-0.4$ comparable with this is seen in the autocorrelation function of the differenced $z$ signal in Supplementary Fig. S3 (b). This estimate is actually within the confidence interval (CI) around $-1/2$, whose size is once more obtainable by the “Fisher law” for the standard deviation of the distribution of sample correlations $1/\sqrt{n-3}$, where $n = 116$ this time is the time series length, which is to be multiplied by about 2 for getting the 95% CI half width, yielding 0.1844. Therefore, we can conclude that no correlation in the $z$ signal could be detected.

The question still is what the error is of the p-value of the original MK test due to the possible small correlations, i.e., to the violation of the test’s assumption. The state-of-the-art answer to this question is given by the modified MK test, namely, that the error is rather small. We mention that the implementation of the modified MK test that we used employs only linear detrending, which is not correct. Nevertheless, with two different linear detrending schemes,
one with the usual least-squares method and another fitting method due to Sen (1968), very similar p-values are found: $1.8 \times 10^{-3}$ with the former, and $2.4 \times 10^{-4}$ with the latter (cf. $p = 4 \times 10^{-4}$ for the original MK test). This seems consistent with the claim that the incorrect detrending in this case would not introduce an error that would alter the significance of the detection of nonstationarity.

Supplementary Fig. S3. The autocorrelation function of (a) $z_i$ and (b) $(z_i - z_{i-1})$ for the MPI-HE and using the SOI to represent the ENSO (where the index of $z_i$ refers to the data point, i.e., to the year). The horizontal blue lines correspond to the interval outside which the correlation coefficient is different from zero at the significance level of 0.05.


Part IV. Checking the effects of non-Gaussianity in the original data

One condition for the approximate Gaussianity of the sampling distribution of the Fisher-transform $z$ of the correlation coefficient is the Gaussianity of the original data between which the correlation coefficient is computed. The usual method for dealing with non-Gaussian data is to replace them with their rank-based inverse normal transform (Bishara and Hittner, Reducing Bias and Error in the Correlation Coefficient Due to Nonnormality, Educational and Psychological Measurement 75(5), 785-804, 2015) in the given sample set (consisting of the different ensemble members in our case, separately in each year). While the resulting estimate for the correlation coefficient will be biased, it seems reasonable to assume that nonstationarity between different years will not be introduced if it is not present for the correlation coefficient computed from the original data. Therefore, repeating hypothesis testing for the rank-based inverse normal transformed data for the presence of trends according to the methodology of the main text should reveal how much the results are affected by non-Gaussianity.

Supplementary Fig. S4 is very similar to Figs. 3(a) and (b) of the main text, indicating that non-Gaussianity has a negligible effect, especially when compared to the sensitivity to the choice of variables as discussed in Section 4 of the main text.

*Supplementary Fig. S4. Same as Figs. 3(a) and (b) of the main text for rank-based inverse normal transformed variables.*
Part V. Testing stationarity with a $t$-test

Under the negligibility of correlations between consecutive years and of non-Gaussianity in the original variables (supported in Parts III and IV of this Supplementary Material), it would follow from the null-hypothesis of a stationary natural measure that the Fisher-transforms $z$ of the estimates of the hypothetical constant coefficient calculated in the different years would be independently drawn samples from a known distribution: a Gaussian distribution with a standard deviation of $1/\sqrt{N-3}$, where $N$ is the ensemble size. The mean of this distribution is unknown but, of course, fixed in time.

We could thereby test by a Kolmogorov-Smirnov test if the estimates from the full time series or from a subinterval of it (denoting the corresponding p-value by $p_{KS}$) might originate from this known distribution with arbitrary mean. However, we have found (results not shown) that this is not a sensitive test to detect nonstationarity; for any mean value, $p_{KS} > 0.05$ even for the full span of the simulations, both when using Niño 3 and SOI to characterize ENSO.

Nevertheless, this suggests that the assumptions of an unpaired two-sample $t$-test (Fisher, 1936) are not seriously unmet. In particular, the $t$-test can be used to detect nonstationarity by splitting the time series of $z$ e.g. to two halves, and checking if the estimates of $z$ in the first and the second half of the time series may come from populations with the same mean or not ($p_t$). (Note that splitting is done with respect to time, so that we are testing for the stationarity of the time series in the usual, temporal sense.)

Applying this methodology for successively smaller subintervals of the full time series, however, reveals a self-inconsistency as follows. If rejection is found for the two halves, this explicitly informs us about the different distributions in these two halves. If we then take e.g. the first half and carry out the same splitting-and-comparing for its first and second half (quarters of the original time series), it may reveal nonstationarity within this first half of the full time series, indicating that the conditions for the original $t$-test were not met, depriving its p-value of its probabilistic meaning.

Nevertheless, this methodology is still suitable for the rejection of the null-hypothesis that the full time series is stationary. In case it is stationary, the conditions for a $t$-test between two arbitrary subintervals are met. If a small p-value is found in such a $t$-test, it may be so according to the very chance given by that p-value under a true null-hypothesis. The only alternative explanation is a false null-hypothesis even if it implies the inapplicability of the $t$-test. Note on the other hand that false negative results (type II errors) as artifacts are not excluded, one is “protected” from an increased chance of false positive results (type I errors) only.

Numerical results for our geophysical problem are presented in Supplementary Fig. S5. They are to be compared with Supplementary Fig. S6, showing the p-values corresponding to Figs. 3(a) and (b) of the main text. While the presence of artifacts, appearing as a transverse diagonal stratification at least, is rather obvious in Supplementary Fig. S5, the pattern of low p-values follows that of Supplementary Fig. S6 very closely, both on long and short temporal scales.
Supplementary Fig. S5. The \( p_t \) values (color coded: \( p_t < 0.05 \): black, \( p_t < 0.005 \): green, \( p_t < 0.0005 \): blue) calculated in the MPI-HE and MPI-RCP8.5 stitched together for all possible subintervals of the whole time span. ENSO is represented by (a) Niño 3 and by (b) SOI.

Supplementary Fig. S6. Same as Supplementary Fig. S5. for the \( p_{MK} \) values (corresponding to the \( Z_{MK} \) values in Fig. 3 of the main text).