Electronic Supplementary Material for:

Reconsidering the relationship of the El Niño–Southern Oscillation and the Indian monsoon using ensembles in Earth system models

Mátyás Herein[1,2,*], Gábor Drótos[2,3,4], Tamás Bódai[5], Frank Lunkeit[1], Valerio Lucarini[1,5,6]

[1]CEN, Meteorological Institute, University of Hamburg, Hamburg, Germany

[2]MTA–ELTE Theoretical Physics Research Group, and Institute for Theoretical Physics, Eötvös University, Budapest, Hungary

[3]Instituto de Física Interdisciplinar y Sistemas Complejos, CSIC-UIB, Palma de Mallorca, Spain

[4]Max-Planck-Institut für Meteorologie, Hamburg, Germany

[5]Centre for the Mathematics of the Planet Earth, Department of Mathematics and Statistics, University of Reading, Reading, UK

[6]Walker Institute for Climate System Research, University of Reading, Reading, UK

*hereinm@gmail.com

Part I. 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):

$$SOI = 10\,\frac{p_{diff}(t) - \overline{p_{diff}(t)}}{\sqrt{\overline{p_{diff}(t)^2} - \overline{p_{diff}(t)}^2}}\,. \qquad \text{(S1)}$$

Here p_{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_{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).

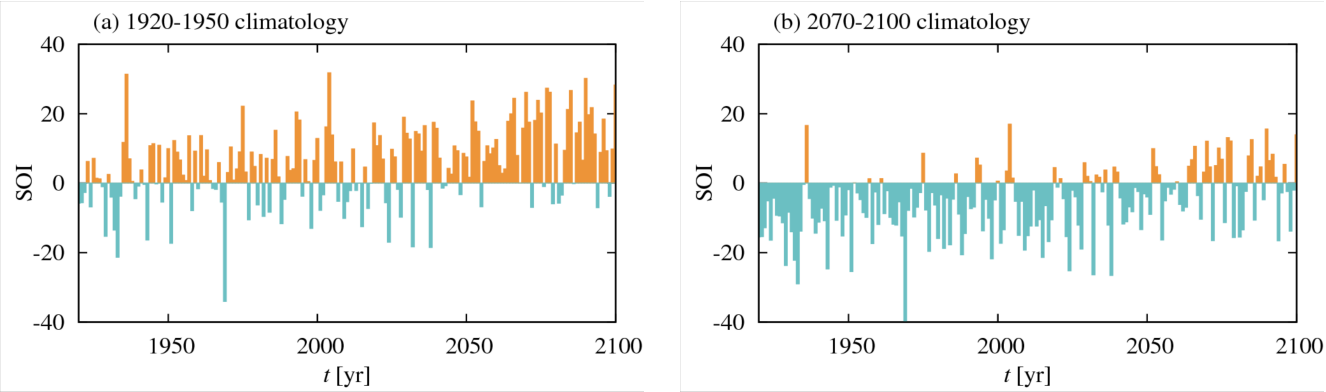
Although there exist sophisticated methods for removing trends from time series, they can resolve the problem only approximately without an a priori knowledge of what should be identified as a trend (i.e., how the real expectation value of a given quantity evolves in time). Furthermore, the experience 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 normalize the sea-level pressures first and take the difference afterwards (e.g. Trenberth, 1976; 1984).

All conceptual problems are resolved, however, by a new, snapshot-based SOI (which we denote by SOI_{E}):

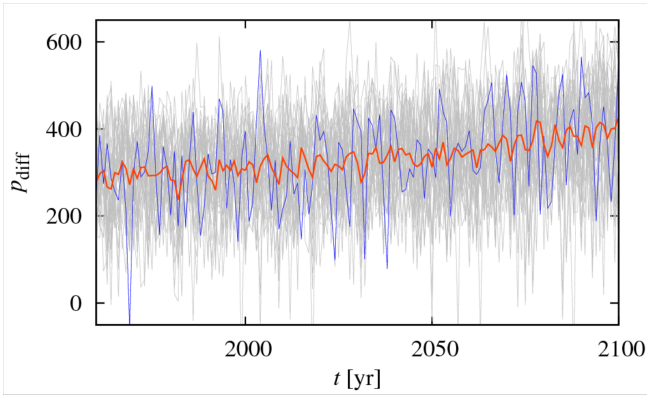
$$SOI_E = 10\,\frac{p_{diff}(t) - \langle p_{diff}(t)\rangle}{\sqrt{\langle p_{diff}(t)^2\rangle - \langle p_{diff}(t)\rangle^2}}\,, \qquad \text{(S2)}$$

where <...> 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_{diff} of one given realization (note that this is the modeling equivalent of an instrumental record) from the *expectation value* of p_{diff}, normalized by the standard deviation. This is so in *any* year, as a consequence of which a natural detrending is provided: the climatic mean of SOI_{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 snapshot-based 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, and Section 5 in the main text.



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 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. p_{diff} of (S1) has been calculated according to Appendix B.



Supplementary Fig. S2. The November sea level pressure difference (p_{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 increase (a "hockey stick") after the year 2050. p_{diff} has been calculated according to Appendix B.

Part II. 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 MPI-HE, since this is the only one in which we obtained p-values below 0.05, resulting in the rejection of the null hypothesis.

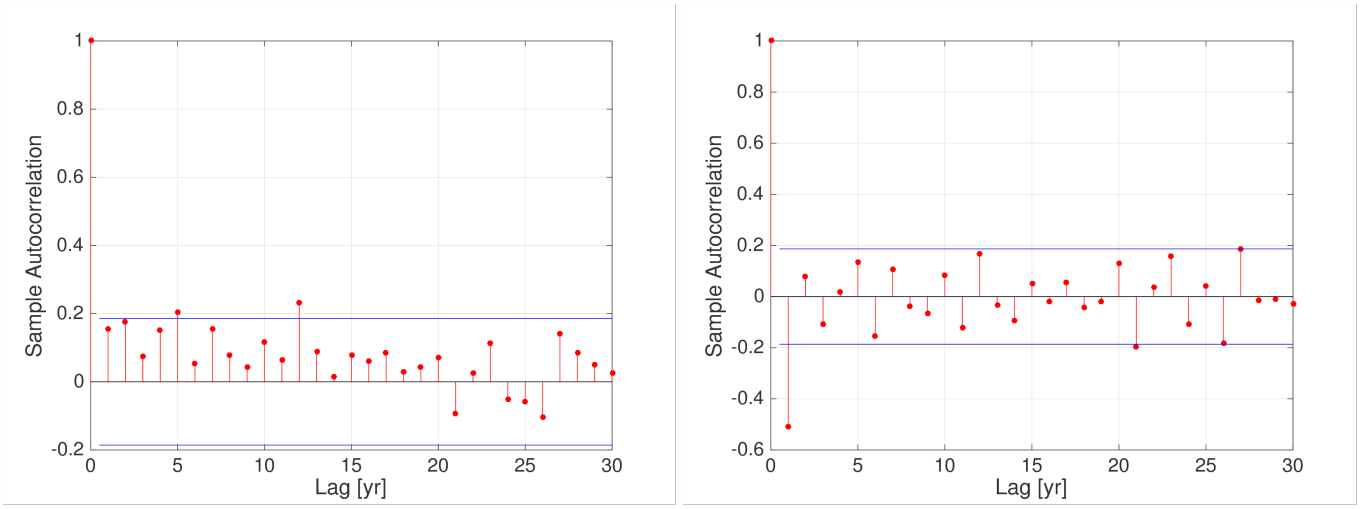
Obtaining very similar p-values by the original and the modified tests is clearly to do with very weak correlations in the time series, if any. This is indicated by a straightforward calculation of the temporal autocorrelation function displayed in Supplementary Fig. S3a. For this we employed the Matlab 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 shape of the investigated distribution (a Gaussian) and its standard deviation ($1/\sqrt{(N-3)}$, where N is the ensemble 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.

Clearly, 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 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 0.1037, while the true value for a stationary z, calculated as $1/\sqrt{(N-3)}$, would be 0.1015, which is very close to the previous value.

Furthermore, we can obtain a kind of a linear estimate for the true signal by fitting a linear trend as a function of the radiative forcing Q (see Supplementary Table S4 and the related discussion in part VII of the Supplementary Material). In this signal, we can take the numerical absolute difference between the consecutive data points (i.e., years). The maximal value of this difference along the time series is 0.0680, and it is below 0.01 for the majority of the years. As these differences are considerably smaller than the above-mentioned values for the standard deviation, we obtain a further support for the assumption that the incremental changes originating from the numerical noise dominate the trend in the original time series of z.

It can be shown easily that the differencing of an uncorrelated stationary signal leads to a $-1/2$ lag-1 autocorrelation. Since quite precisely this value is seen in the autocorrelation function of the differenced z signal in Supplementary Fig. S3b, we can conclude that any undetected correlation in the z signal should be rather small.

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: 4.1×10^{-5} with the former, and 1.05×10^{-4} with the latter (cf. $p = 2.1\times10^{-5}$ for the original MK test, already given Table 1). 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 (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 III. Effects of the ensemble size

To check whether the ability to pose stronger statements for the MPI-HE in Table 1 originates from the larger size of this ensemble, we take 10000 examples of smaller subsets of the MPI-HE that are of the same size as the other three ensembles (77, 68, and and 35 members, respectively), and calculate the proportions q (a Monte Carlo-type probability $\mathcal{P}$) in which stationarity is rejected according to p_{t12} and p_{MK0}. Given in Supplementary Table S1, the high proportions for the 77-member and the 68-member subsets of the MPI-HE suggest that failing to reject stationarity in the MPI-RCP8.5E or the MPI-1pctE is not due to their smaller size. The more moderate proportions for the size of 35 members leaves the same question open for the CESM-LE.

	$q = \mathcal{P}(p_{t12} < 0.05)$	$q = \mathcal{P}(p_{MK0} < 0.05)$
77 members	0.99	0.999
68 members	0.96	0.99
35 members	0.55	0.69

Supplementary Table S1. The proportion q in 10000 subsets of the MPI-HE of given size in which $p_{t12} < 0.05$ *and* $p_{MK0} < 0.05$*, respectively.*

Part IV. The forced response of the correlation coefficient as obtained by evaluation over time

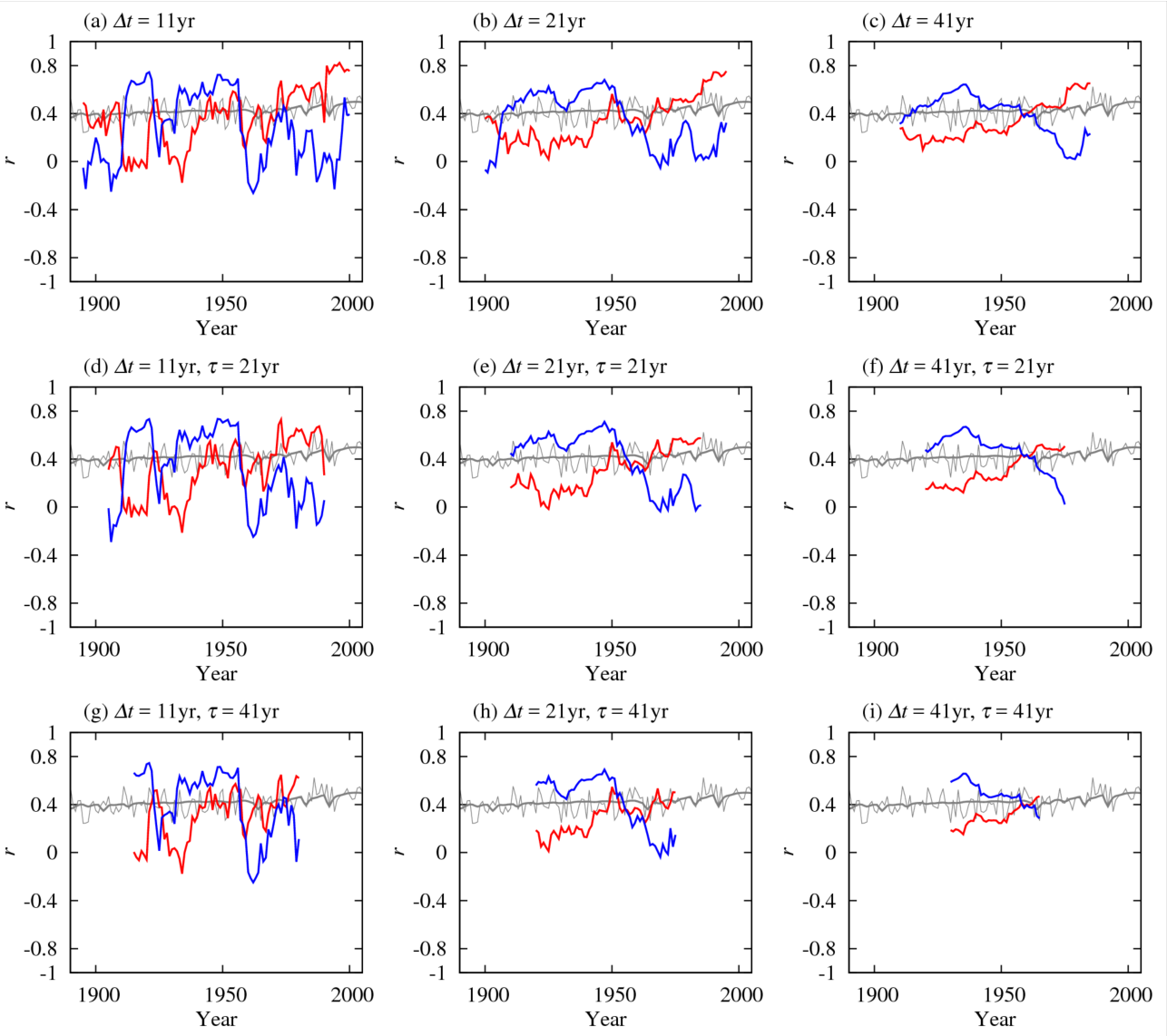
We shall illustrate here that the traditional technique for evaluating correlation coefficients and for investigating their time evolution can lead to strongly misleading results in our case. The traditional technique takes a single realization, and calculates the correlation coefficient with respect to time within some given time interval, a window, of length Δt. The time evolution of the correlation coefficient is obtained in this case by moving (sliding) this window along the time series. Since forced trends can obscure the relationship between the fluctuations, some kind of detrending of the two time series to be compared is usually needed. For illustrative purposes, we choose here one of the simplest detrending techniques: we subtract a moving average, calculated within a time window of length τ, from the original time series.

We calculate the time evolution of the JJA correlation coefficient for two different members of the MPI-HE using the above-described technique, and we compare several values of the freely chosen parameters Δt and τ (including a calculation without detrending, too). In Supplementary Fig. S4, we compare the results to each other, to the actually observed time evolution (the forced response) in the ensemble, and to an estimate obtained by linearly regressing the Fisher-transform z of the correlation coefficient to the radiative forcing Q (see part VII of the Supplementary Material, and Supplementary Table S4 in particular). It is obvious that the traditionally evaluated signals exhibit very little similarities with the correctly evaluated one and with the linear regression. In particular, the fluctuations are typically much larger, and long periods exhibiting apparent, false trends can be seen. This particular example is not sensitive to detrending, but the choice for the time window Δt, over which the correlation coefficient is evaluated, is important: with increasing Δt, the fluctuations become smaller, but the length of the periods with false trends increases (and, as a result, the slope of these false trends decreases). Nevertheless, the main character of the signal in a particular realization is similar for different values of Δt.

Note that both realizations can serve as an example for what can be instrumentally recorded on a planet whose climate system is described by the MPI-ESM, and which is subject to the historical forcing. It is then striking to see how different time evolution ("forced response") can be obtained for the correlation coefficient in our two examples. On our hypothetical planet, climatologists in realization 1 (red in Supplementary Fig. S4) would conclude that the teleconnection between the ENSO and the Indian summer monsoon underwent a very strong strengthening in the 20th century, from nearly negligible to very significant. On the same planet with the same forcing, climatologists in realization 3 would identify, from generally high values, a strong drop in the 1960s in the strength of the teleconnection, from which the strength can hardly "recover". This strong dependence on the particular realization (note that all realizations are equally probable) illustrates that it is very hard (or maybe impossible) to draw conclusions about the forced response of the strength of teleconnections to greenhouse-gas forcing based on a single realization. For a more detailed analysis in an intermediate-complexity climate model, see Herein et al. (2017).

The instrumental observation shown by Yun and Timmermann (2018) — without performing a formal statistical test — can be seen to pass as a possible realization of the MPI-ESM, showing a large variability throughout the 20th century. The instrumental observation shown by Krishna Kumar et al. (1999) has a very different character: it is a "hockey stick", with considerably less variability before 1980. Because of this characteristic did the authors suggest that the decline in the teleconnection could be an emerging signal of forced response. If this feature were to be credited as objective, i.e., not an artifact, then it would prompt that the MPI-ESM is missing a major effect in the ENSO-Indian monsoon teleconnection.

It is actually an open question if the fluctuations ("modulations") of the correlation coefficient evaluated with respect to time in a single realization are related to some low-frequency mode of internal variability (cf. Section 6 of the main text). Even in this case, these fluctuations can be considered artificial from the point of view of a forced response, since they do not imply any changes in the "true" correlation coefficient, the one that fully characterizes internal variability (see the mentioned Section). However, the strong dependence on Δt suggests that at least the observed characteristics of the trend-like fluctuations in our example do not have such an origin. In particular, they are presumably the manifestation of the effects described in Wunsch (1999), Gershunov et al. (2000) and Yun and Timmermann (2018).



Supplementary Fig. S4. The time evolution of the JJA correlation coefficient r, plotted as a function of the time t, between the sea level pressure difference p_{diff} and the Northern Indian precipitation P, in two realizations of the MPI-HE. The red and the blue line correspond to realizations 1 and 3, respectively. For comparison, the ensemble result and a linear regression (see see part VII of the Supplementary Material, and Supplementary Table S4 in particular) are also included as a thin and a thick gray line, respectively. In the different panels, different window lengths for the evaluation of the correlation coefficient (Δt) and for the detrending (τ) are considered (in the upper row, no detrending is applied). See text for details.

Part V. If radiative forcing were dynamical forcing with an instantaneous linear response

After the surprising result that response is detectable only in the radiatively most weakly forced setup (i.e., in the MPI-HE), we investigate its implication for the dynamical role of the radiative forcing Q. Although our investigation works with rather naive assumptions, the results will be indicative of some general conclusion.

In order to have an impression of what the ability of detecting a trend means, we check the sensitivity of our test of p_{MK0} to the presence of a particular kind of a "mostly increasing trend" in the strength of the teleconnection. We define this kind of signal as a *linear* increasing relation between the Fisher-transform z of the correlation coefficient and the radiative forcing Q (that is, *not* the time t). We assume this relation to hold at any time instant, i.e., that the response to radiative forcing is instantaneous, without any delay. We pose our assumption for the Fisher-transform z of the correlation coefficient instead of posing it for the correlation coefficient r itself, because the value of the former (i.e., the area hyperbolic tangent of the latter) is unbounded, so that possible deviations from linearity that arise from a bounded range can be excluded. Note that a strong implication of our assumption is that the radiative forcing Q can serve as the dynamical forcing which the system is subject to.

By checking the sensitivity of our test to the kind of signal as defined above (not to be confused with climate sensitivity, i.e., the sensitivity of some statistics, or the entire distribution (Chekroun et al., 2011), of a variable of the climate system with respect to a parameter), we mean that we look for the *weakest* such relation that results in a time series in which p_{MK0} detects a trend at a significance level of 0.05 with a given probability $q = \mathcal{P}(p_{MK0} < 0.05)$. We take the actual radiative forcing scenarios and ensemble sizes, and assume the same temporal mean for the Fisher-transform z as the one observed in the actual ensembles. See part VI of the Supplementary Material for the details of our Monte Carlo algorithm, which is based on 100000 random time series for the circumstances of each ensemble, and which estimates the probability $q = \mathcal{P}(p_{MK0} < 0.05)$ as the corresponding proportion among these 100000 time series.

The slope χ between the Fisher-transform z of the correlation coefficient and the radiative forcing Q assuming an instantaneous linear relation between these two variables (the latter of which represents the dynamical forcing under the given assumption) is, in fact, the static susceptibility (the Fourier transform of the response function taken at zero frequency) of the former variable with respect to the latter one in the terminology of nonequilibrium statistical mechanics (Kubo et al., 1991); see Ruelle (2009) for susceptibilities in dynamical systems. Supplementary Table S2 gives the results for the sensitivities of our test of p_{MK0} in the form of the smallest slopes χ that would *just* be detected with two given probabilities q: $q = 0.50$ gives the turning point to a more probable detection of the trend than not, and $q = 0.95$ gives a trend that is "almost certainly" detected. In Supplementary Table S2, the values of the correlation coefficient r that would be present at the beginning and the end of the given simulations with the obtained slopes are also shown.

The results in Supplementary Table S2 indicate that our hypothesis test is, in terms of the slope, much less sensitive in the MPI-HE than in the other three ensembles (i.e., hypothetical nonstationarities of the time series associated with small slopes χ are not detectable in the MPI-HE, only those in association with steep ones), while these other three ensembles are characterized by sensitivities similar to each other. The former finding is a natural consequence of the particular ranges of the radiative forcing Q in the particular ensembles: this range is small in the MPI-HE (see Fig. 1), therefore, a steep slope χ (steeper by a factor of 3) is needed to be present to detect a similar "mostly increasing trend" of the time series of z as in the other ensembles. Note, however, the counterintuitive nature of the fact that the MPI-HE is the least sensitive ensemble in terms of the slope (i.e., unlike in the other ensembles, small slopes cannot be detected), yet it is the only ensemble in which we could actually detect nonstationarity (a nonzero slope). The sensitivity in terms of an other measure, in that of the change in the correlation coefficient r from the beginning to the end of the simulations, i.e., in terms of the detectable signal in r, is similar in all ensembles.

	$q = \mathcal{P}(p_{MK0} < 0.05)$	χ [1/(Wm^{-2})]	r at the beginning	r at the end
MPI-HE	0.50	0.046	0.39	0.47
	0.95	0.086	0.36	0.51
MPI-RCP8.5E	0.50	0.013	0.41	0.48
	0.95	0.024	0.38	0.50
MPI-1pctE	0.50	0.014	0.41	0.47
	0.95	0.026	0.38	0.50
CESM-LE	0.50	0.013	0.07	0.16
	0.95	0.024	0.04	0.21

Supplementary Table S2. The slope χ of the weakest linear increasing relation between the Fisher-transform z of the correlation coefficient and the radiative forcing Q that is detected by p_{MK0} at the significance level of 0.05 with a probability q, under the circumstances of the given ensembles. The corresponding values of the correlation coefficient r are given for the beginning and the end of the simulations.

	$q = \mathcal{P}(p_{t12} < 0.05)$	χ [1/(Wm^{-2})]	r at the beginning	r at the end
MPI-HE	0.50	0.063	0.38	0.49
	0.95	0.121	0.33	0.54
MPI-RCP8.5E	0.50	0.014	0.41	0.48
	0.95	0.027	0.38	0.51
MPI-1pctE	0.50	0.016	0.40	0.48
	0.95	0.030	0.37	0.51
CESM-LE	0.50	0.014	0.06	0.17
	0.95	0.026	0.03	0.22

Supplementary Table S3. The slope χ of the weakest linear increasing relation between the Fisher-transform z of the correlation coefficient and the radiative forcing Q that is detected by p_{t12} at the significance level of 0.05 with a probability q, under the circumstances of the given ensembles. The corresponding values of the correlation coefficient r are given for the beginning and the end of the simulations.

The same investigation of the sensitivity has been carried out for the test of p_{t12} as well. The results, given in Supplementary Table S3, lead to the same conclusions as for the test of p_{MK0}. However, the quantitative values have to be treated with caution, since we explicitly deviate here from the assumptions of the *t*-test (see Appendix C) by using a linear relation between the Fisher-transform *z* of the correlation coefficient and the ever-changing forcing.

To sum up, we actually detected nonstationarity in the MPI-HE, in which a hypothetical nonstationarity is “detectable” only if it is associated with a slope χ 3 times steeper than those that make nonstationarity “detectable” in the other ensembles. At the same time, we could not reject stationarity in the other ensembles. This means that, in terms of a linear and instantaneous

relation, the strength of the response to radiative forcing, i.e., the static susceptibility χ, is estimated to be at least 3 times larger in the MPI-HE than in any of the other three ensembles. For the direct estimation of these susceptibilities, see part VII of the Supplementary Material.

Note that our assumption of a linear response corresponds to the existence of a *single* value of the static susceptibility χ. Our results indicate that this can be true sectionwise at most. A sectionwise linearity with slopes different by a factor of 3, especially for the forcings presented in Fig. 1, does not seem to be plausible. That is, some of our assumptions must be grossly wrong, as we discuss further in Section 6 of the main text.

Part VI. Estimating the sensitivity of the Mann-Kendall test in the particular ensembles

Besides the p_{MK0} value according to which we reject or not stationarity, it is also important to know how strong nonstationarity needs to be present for a rejection — this is what we regard as the sensitivity of the Mann-Kendall test. This sensitivity, of course, depends on the particular choice of the significance level p_{sig} for rejection.

Without additional constraints for the nonstationary signal, the sensitivity cannot be determined. With regards to the underlying physical process, we determine the sensitivity to the presence of an instantaneous, linear increasing relation between the Fisher-transform z of the correlation coefficient and the radiative forcing Q (i.e., not the time t).

By choosing an instantaneous relation we neglect the delay that is certainly present (Herein et al., 2016); this delay can be, however, assumed to be small compared to the time scale of the changes in the entire simulations. Unfortunately, we would be able to estimate the delay only if we knew the precise time series of the Fisher-transform z. It is even more important to recall from Section 2 of the main text that the radiative forcing Q is not the dynamical forcing which the system is subject to. By assuming a functional relationship between a variable (the Fisher-transform z in our case) and the radiative forcing Q, we implicitly also assume that the latter can serve as the dynamical forcing. This can be regarded as an approximation, which can prove to be invalid if the data does not fit well to the assumed functional relationship. At least for the MPI-HE, a reasonable fit can be found (cf. Supplementary Table S4 and Supplementary Fig. S5), but note that this does not imply that the assumption or approximation is principally correct.

Assuming the linear relation defined above, we look for the weakest slope that results, under the imposed forcing of each simulation, in a time series that is rejected by the Mann-Kendall test to miss any monotonic trend (in what follows, we shall call this as the weakest nonstationarity that is 'detectable' or 'detected'). Note that the radiative forcing is different in each ensemble simulation analyzed in our study, which is one reason why we have to carry out the estimation separately for each ensemble. Although there is a forcing scenario, the historical one, in which the radiative forcing Q is not perfectly monotonic in time, a rejection still implies in this case that the corresponding, nonmonotonic time series of the Fisher-transform z cannot be stationary.

The linear dependence of the Fisher-transform z on the forcing with a given slope still does not determine the observed time series of z. Instead, each data point (corresponding to one particular year) in this time series is a sample drawn from a Gaussian distribution, the mean of which is the actual Fisher-transform z in the given year, and the standard deviation of which is $1/\sqrt{(N-3)}$, where N is the ensemble size (Fisher, 1936). (Note that this implies that data points of different years are drawn from different distributions, which differ in their mean, but not in their shape and standard deviation.) Because of the stochastic nature of the generation of the series, the weakest nonstationarity that is detectable cannot be determined definitely. For this reason, we proceed as follows, separately for each ensemble.

We assume that a particular slope is present, and we generate 100000 different time series according to the time-dependent distribution described in the previous paragraph. Among these 100000 different time series, the proportion q of those in which rejection occurs is our Monte Carlo estimate for the probability $\mathcal{P}$ of detecting the nonstationarity. By varying the slope with successive approximation, we find the slope that corresponds to a given, prescribed detection probability $q = \mathcal{P}(p < p_{sig})$. Two intuitive choices for q are $q = 0.50$, which gives the turning point to a more probable detection of the trend than not, and $q = 0.95$, which gives a trend that is “almost certainly” detected. The slope obtained this way is what we regard to characterize the sensitivity of the Mann-Kendall test to our assumed form of nonstationarity.

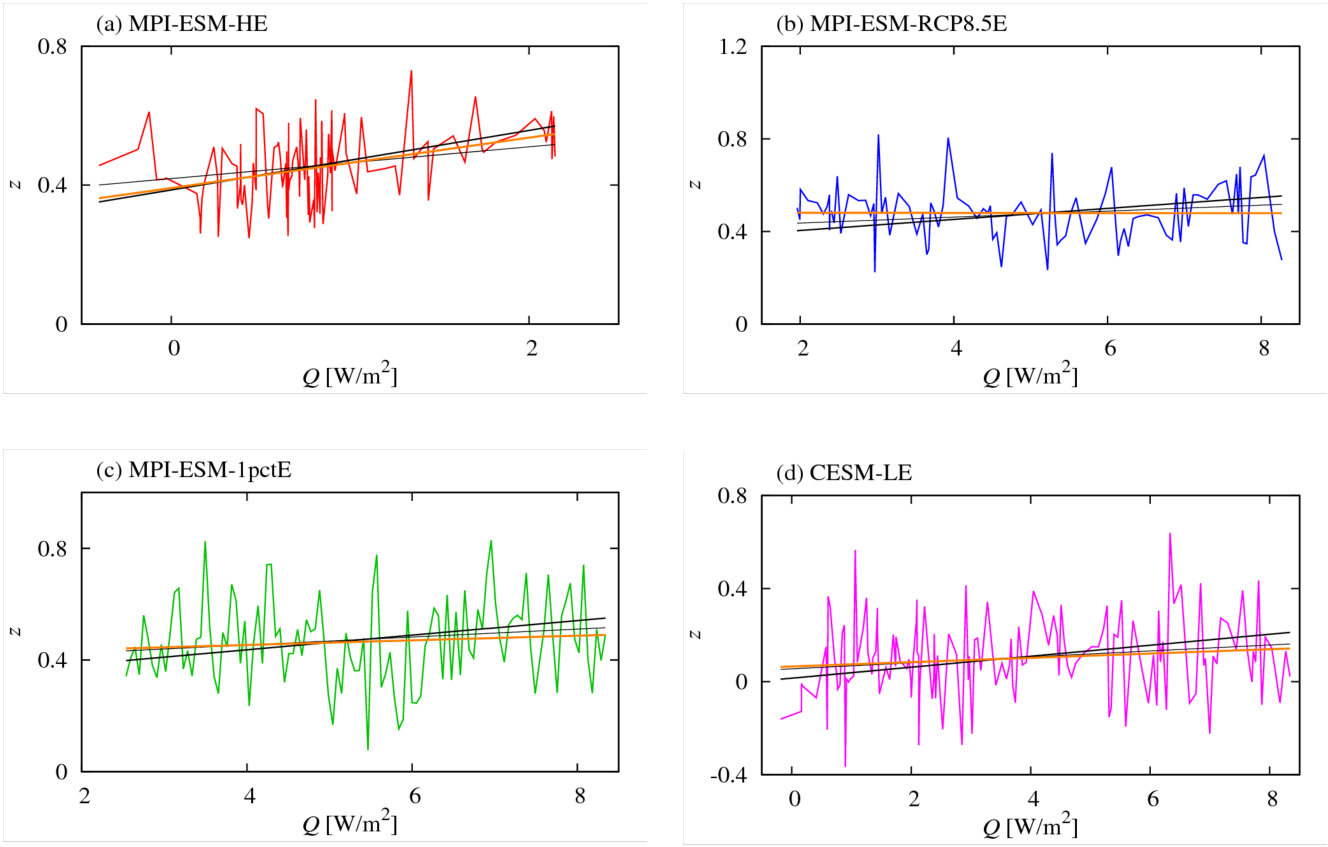
Part VII. A direct estimation of the susceptibilities

Here we directly estimate the susceptibilities χ discussed in part V of the Supplementary Material. If we assume a linear response in terms of the Fisher-transform z of the correlation coefficient as a function of the radiative forcing Q, the maximum likelihood estimate is given by the least squares linear regression (Press, 2007), since the errors of the data points, as mentioned, come from a Gaussian distribution with the same standard deviation (Fisher, 1936), and were found to be independent from each other. Supplementary Table S4 gives the parameters, with their uncertainty, of the numerically fitted lines of the form of $z = z_0 + \chi Q$. It becomes obvious that we cannot conclude about any pronounced relationship in the cases of the MPI-RCP8.5E, the MPI-1pctE, and the CESM-LE, while we find a well-fitting, positive-sloped line in the MPI-HE. This finding is also confirmed by the direct visual observation of the Fisher-transform z of the correlation coefficient as a function of the radiative forcing Q, shown in Supplementary Fig. S5. (Note in this figure that the regression line, i.e., the most likely linear relationship, is always less steep than what would be detectable by p_{MK0} more probably than not ($q = 0.50$), except for the MPI-HE, in harmony with our finding that we can reject stationarity only for the MPI-HE.)

Since the relationship between the Fisher-transform z of the correlation coefficient and the radiative forcing Q has been found to be approximately linear, and Q increases mostly in the second half of the 20th century within the time span of the MPI-HE (see Fig. 1), we conclude that the increase in the strength of the teleconnection between the ENSO and the Indian monsoon is also concentrated to this period. This is confirmed by Fig. 4.

	χ [1/(Wm^{-2})]	z_0
MPI-HE	0.073 ± 0.016	0.392 ± 0.016
MPI-RCP8.5E	-0.0004 ± 0.006	0.482 ± 0.033
MPI-1pctE	0.008 ± 0.008	0.421 ± 0.048
CESM-LE	0.009 ± 0.006	0.065 ± 0.027

Supplementary Table S4. The parameters, with their standard errors, of the least squares linear regression $z = z_0 + \chi Q$ between the Fisher-transform z of the correlation coefficient and the radiative forcing Q, in the different ensembles.



Supplementary Fig. S5. The Fisher-transform z of the correlation coefficient as a function of the radiative forcing Q, plotted with lines connecting neighboring datapoints. Thin and thick black lines correspond to the weakest linear increasing relations, from Supplementary Table S2, that would be detectable by p_{MK0} *at a significance level of 0.05 with a probability of q = 0.50 and q = 0.95, respectively. The thick orange line is the least squares linear regression from Supplementary Table S4. The different panels consider different ensembles.*

Part VIII. The climatic mean's forced response as obtained by temporal averaging

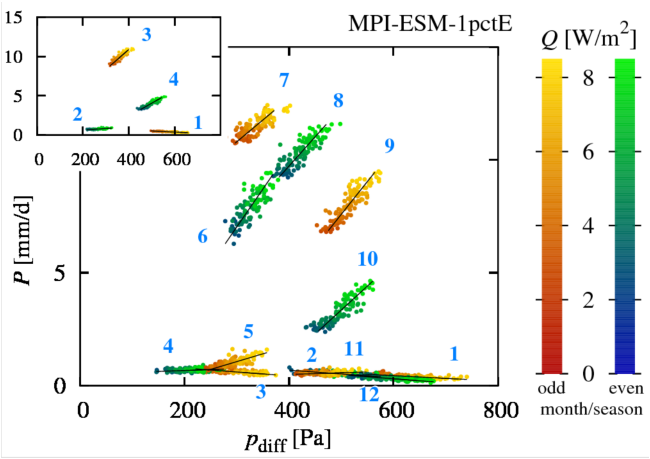
We shall check here what is obtained numerically in the phase-space projection chosen in our paper when the climatic mean is calculated by the traditional technique, which takes a temporal average for one time series (corresponding to a single member in the ensemble). For a shifting time series, a moving average needs to be taken to obtain the time evolution of the climatic mean.

For this investigation, we shall consider an arbitrarily chosen member of the MPI-1pctE. For reference, the ensemble result (the counterpart of Fig. 5a or 5b) is given in Supplementary Fig. S6.

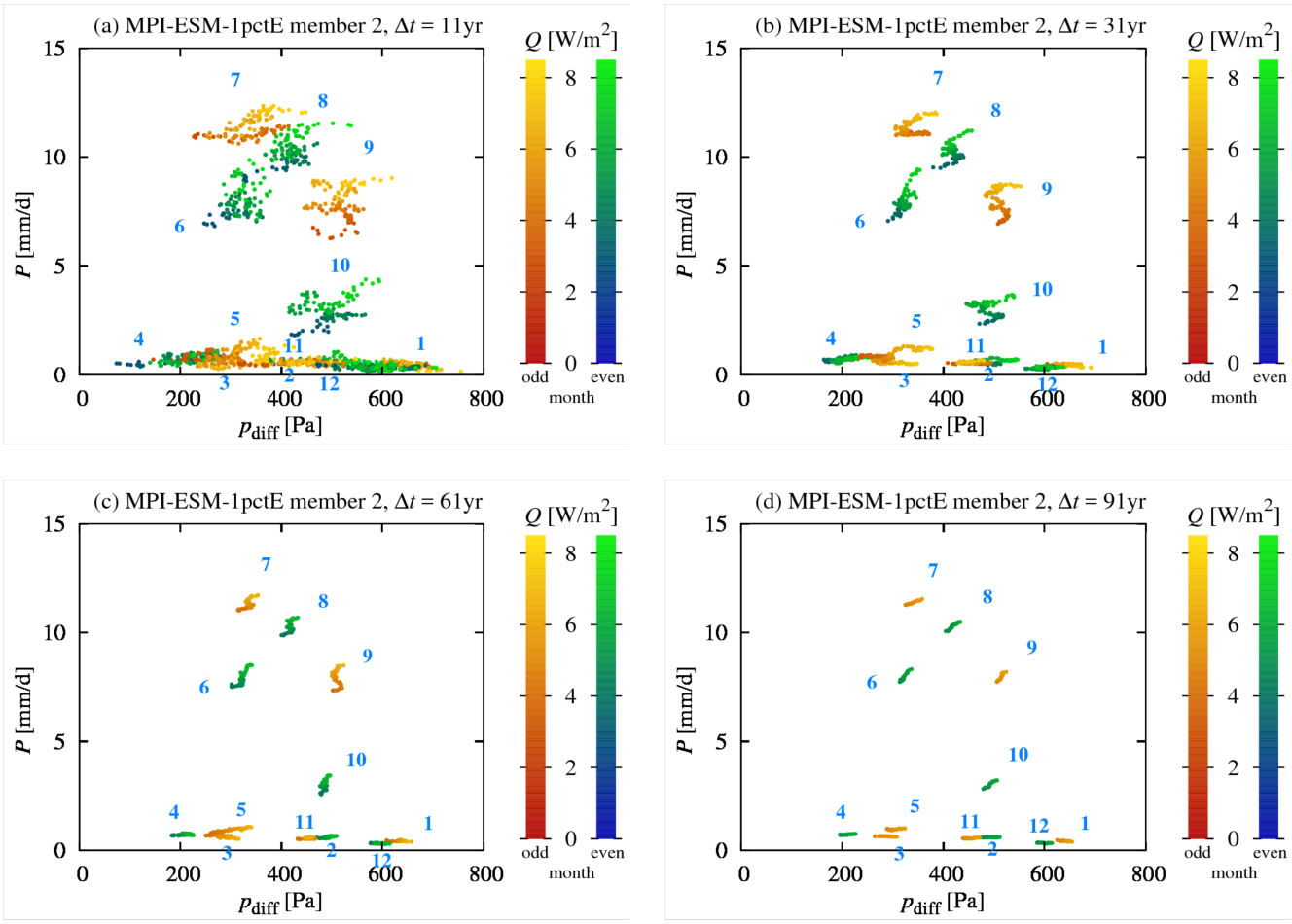
Supplementary Fig. S7 presents the numerical results obtained with the traditional technique for different window lengths τ. It is obvious for τ = 11yr (Supplementary Fig. S7a) that the internal variability is so strong that each month appears as a cloud of points without a prominent structure. Nevertheless, the clouds are elongated to some extent, and this elongation might be thought to represent the linear behavior identified in Fig. 5a and similarly present in Supplementary Fig. S6. However, due to the inability of separating the effect of the internal variability from the forced response, we could not find any well-grounded method to fit lines to the clouds of points. A visual inspection may find the main direction of the elongation to be less steep in each month than the slope of the corresponding line in Supplementary Fig. S6. The reason for this is the much larger internal variability in p_{diff} than in P. We thus learn that variable-dependent internal variability may introduce systematic errors into the interpretation of the time evolution represented in the corresponding phase-space projection.

One might try to filter out internal variability using longer windows. In Supplementary Figs. S7b-d, the sets of points of individual months become less fuzzy indeed, they appear like curves instead of clouds. Since the ratio between the extension due to a real trend (the real response) and that due to internal variability (a signal to noise ratio) increases, the overall directions of the particular sets of points tend to get closer to the correct ones. Locally, however, we are facing very strong false trends for τ = 31yr and 61yr (Supplementary Figs. S7b-c): the local tangent of the curves deviates from the correct one (i.e., from the direction of the lines in Supplementary Fig. S6) by more than 90 degrees for sections corresponding to several decades. This means trends that are incorrect in their sign for the individual variables. Such false trends are characteristic to moving averages (Wunsch, 1999), and similar effects have been pointed out by Drótos et al. (2015) for one variable in the same context of climatic averages.

For τ = 91yr (Supplementary Fig. S7d), the false trends disappear, but the directions of the curves are still not reliable (note the spread between the different months from June to October, which is not present in Supplementary Fig. S6). At the same time, we are close here to reaching the upper bound for τ imposed by the length of the original data. To sum it up, we have found that temporal averaging for calculating responses in climatic means fails for the simulations considered in this paper. As for more extended data sets for which τ can be further increased, nonlinearities might appear in the response, for which temporal averaging introduces biases (Herein et al., 2016; Drótos et al., 2016). Therefore, temporal averaging cannot be applied without additional considerations in such cases either.



Supplementary Fig. S6. Same as Fig. 5a or 5b for the MPI-1pctE.



Supplementary Fig. S7. The traditional climatic mean (obtained as a temporal average) in the sea level pressure difference p_{diff} and the Northern Indian precipitation P. All different months are plotted, see the numbering (1-12: January-December). For a given month or season, each data point represents a particular year, on which the time window for averaging is centered. The different years are colored according to the color scales on the right. The different panels correspond to different window lengths τ, as indicated.