Ignoring internal variability can lead to misleading conclusions on model fidelity

Shipra Jain[1], Adam A. Scaife[2,3], Theodore G. Shepherd[4,5], Clara Deser[6], Nick Dunstone[2], Gavin A. Schmidt[7], Kevin Trenberth[6], Thea Turkington[1]

[1]Centre for Climate Research Singapore (CCRS), Singapore

[2]Met Office Hadley Centre, United Kingdom

[3]University of Exeter, United Kingdom

[4]Department of Meteorology, University of Reading, United Kingdom

[5]Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany

[6]National Center for Atmospheric Research, United States of America

[7]NASA Goddard Institute for Space Studies, United States of America

Abstract

Benchmarking climate model simulations against observations of the climate is core to the process of building realistic climate models and developing accurate future projections. However, in many cases, models do not match historical observations, particularly on regional scales. If there is a mismatch between modeled and observed climate features, should we necessarily conclude that our models are deficient? Using several illustrative examples, we emphasize that internal variability can easily lead to marked differences between the basic features of the model and observed climate, even when decades of model and observed data are available. This can appear as an apparent failure of models to capture regional trends or changes in global teleconnections, or simulation of extreme events. Despite a large body of literature on the impact of internal variability on climate, this acknowledgment has not yet penetrated many model evaluation activities, particularly for regional climate. We emphasize that using a single or small ensemble of simulations to conclude that a climate model is in error can lead to premature conclusions on model fidelity. A large ensemble of multidecadal simulations is therefore needed to properly sample internal climate variability in order to robustly identify model deficiencies and convincingly demonstrate progress between generations of climate models.

Keywords: Chaotic internal variability, climate model evaluation, future projections, large ensembles.

1. Introduction

Climate models are remarkably successful in reproducing many earth-system phenomena such as atmospheric jet streams, oceanic currents, monsoons, the El Niño Southern Oscillation (ENSO), the North Atlantic Oscillation (NAO), and the response of global climate to external forcing (Broecker, 1975; Hansen et al., 1981, Hausfather et al., 2020). From their basis in the Navier-Stokes equations of fluid dynamics, even extreme events like heatwaves and cold snaps, floods and droughts, cyclones, and storms all appear spontaneously in climate model simulations. In some cases, models also warn us of more intense extreme events than we have yet experienced but which could plausibly occur at any time in the current climate (Jain et al., 2020; McKinnon and Deser, 2021; Thompson et al., 2017; Kent et al., 2022). Scientists also use climate models to understand the physical mechanisms behind past changes in climate (e.g., Mitchell et al., 1995), to understand and predict extreme events (e.g., Hardiman et al., 2020), to project future climate change, and to inform governments and policymakers about the impacts of climate change (IPCC AR6 WG I report, 2021).

To gain confidence in model projections, the fidelity of models is assessed by comparing model simulations of the historical climate with observations. This trial of models using observations is core to identifying current model deficiencies and prioritizing areas for further development (e.g., Flato et al., 2013). The scientific literature presents many examples where climate models apparently fail to reproduce even basic aspects of observed regional climate such as trends in regional rainfall amount and temperature (e.g., Mishra et al., 2018; Mitchell et al., 2013; Saha et al., 2014), multidecadal changes in atmospheric circulation (Gillett et al., 2003) and climatology (Seidel et al., 2008), the frequency or magnitude of extreme events (Tyrrell et al., 2022, Slingo et al., 2022), global teleconnections (Butler and Polvani, 2011), interaction between different modes of climate variability (Martin et al., 2021) or external forcing effects (Schmidt et al., 2014; Legrande et al., 2016). However, even in the case of perfect models, perfect boundary conditions, and perfect observations, a lack of agreement between the modeled and observed climate can still arise simply due to chaotic internal variability (Tebaldi and Knutti, 2007; Deser et al., 2012, 2016, 2018; Mitchell et al., 2013; Jain and Scaife, 2022; Fasullo et al., 2020, McKinnon and Simpson, 2022).

Due to the inevitable presence of internal variability, each realization of climate, in both observed and models, represents only a possible state (Lorenz 1963, 1969). When evaluating models against observations, this issue (known as sampling variability) is supposed to be handled through statistical tests, but the power of those tests relies heavily on the assumed chance process generating internal variability. This leads to a misinterpretation of significance tests, which is widespread in climate science, leading to overly lax criteria (Shepherd, 2021). This problem is only exacerbated if the statistical tests do not take proper account of multidecadal internal variability, which is very difficult to quantify from single or small model ensembles of models, or from observations.

Many of the studies claiming a mismatch between models and observations, including some cited in the IPCC WG1 report (2021), continue to use only a single or a small ensemble of simulations from a given model. These studies generally show that the differences are *statistically significant,* typically for time period considered sufficiently long. Here, we provide three illustrative examples to demonstrate how internal variability cannot be easily ruled out as a cause of commonly reported discrepancies even when decades of observations and model data are available.

2. Do models reproduce observed trends?

Over the latter half of the 20th century, historical observations show a significant reduction in the summer monsoon rainfall over parts of India (Kumar et al., 2010, Ramesh et al., 2014; Saha et al., 2014; Jin and Wang, 2017). However, most historical climate model simulations from different CMIP generations have shown a consistent increase in rainfall during this period and beyond into the 21st century under increasing greenhouse gas forcing (Katzenberger et al., 2021; Saha et al., 2014; Shashikanth et al., 2013).

Several hypotheses have been proposed to explain the apparent mismatch in observed and model-simulated trends in monsoon rainfall over north-central India, including large observational uncertainty (Singh et al., 2019), the recent warming of the Indian Ocean (Roxy et al., 2015), radiative forcing due to aerosols (Monerie et al., 2022) or changes in land-use and land-cover (Paul et al., 2016). In most cases, these factors are found to have systematic effects on the modeled monsoon rainfall. However, it is also possible that internal variability on multidecadal timescales contributes to the mismatch.

We find that the most extreme drying and wetting trends in the Indian Meteorological Department (IMD) observational record for both regional (Fig. 1a) as well as all-India rainfall (Fig. 1b) lie within the range of plausible trends from chaotic internal variability in the current climate. Even more extreme values than have been observed in the IMD record so far are possible, solely due to internal variability, in the absence of any systematic forced effects. Therefore, for this case, internal variability cannot easily be rejected as the cause of the models' apparent failure to capture the observed trend (also see Huang et al., 2020; Sinha et al., 2015).

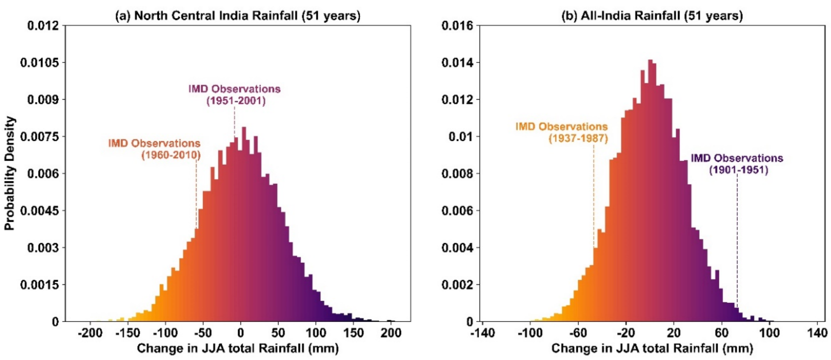

Figure 1 Internal variability in rainfall trends. Change in June-July-August (JJA) total rainfall over a 51-year period from internal variability in a multimodel ensemble (MME) of climate predictions for (a) north-central India (20-28°N, 76-87°E) and (b) all-India (land-only) rainfall. The darker color indicates wetter trends. The dotted lines show the most extreme 51-year trends in the IMD observational record during 1950-2013 (the period for which drying trends were observed over north-central India) and 1901-2013 for all-India. See Methods for more details.

Unprecedented climate extremes are often a manifestation of both internal variability and external forcing. However, in many cases, the internal variability is so large that it can easily negate or greatly overwhelm any forced response in climate trends, even on multidecadal time scales (e.g. Fischer et al., 2013; Jain and Scaife, 2022). For example, in this case, internal variability can be large enough to overwhelm the wetting trend due to greenhouse gas forcing and give temporary drying trends in monsoon rainfall.

Another example where the role of internal variability was ignored is the recent paper by Scafetta (2022), which claimed that no model with a climate sensitivity >3ºC was consistent with observed trends since 1979. However, this claim is based on comparing each model's ensemble mean change with an observational estimate (from ERA5) without taking either observational uncertainty or model internal variability into account. When these elements are included and appropriate statistical tests were employed (following Santer et al, 2008), it was demonstrated that the majority of the models with sufficiently large ensembles, even some of those with high climate sensitivity of 5ºC, are consistent with the observations (Schmidt et al. 2023).

There also exists a selection bias when comparing the most extreme trends in the observational record with model-simulated trends (Rahmstorf et al. 2017). Inspection of observational time series, with no a priori reason to select the particular period of most extreme trends, followed by comparison with model-simulated trends *for the same time period* introduces a selection bias and the impression that models fail to produce observed trends. In many cases, those extreme trends in regional climate can appear at any time in the model simulation and not necessarily during the observed period, irrespective of any forced changes which are often smaller than the internal variability for short- or medium-term periods. It is therefore very difficult to argue that models cannot reproduce observed extremes by pre-selecting extreme periods in the observations and then testing models for the same historical period. Significance testing of these extreme trends is difficult in such cases and simple tests based on a single or limited ensembles of simulations to reject the null hypothesis of internal variability in favor of a model error are often invalid (Eade et al., 2021).

Whilst the Indian rainfall trend presented here is only one example, many other cases of models apparently failing to reproduce observed trends for other regions and variables also exist (e.g. Gillett et al., 2003, Seidel et al., 2008, Swart et al. 2015, Deser et al., 2020, 2023). Therefore, it is necessary to re-examine such cases and carefully rule out internal variability as a cause of apparent mismatch between observed and modeled trends to robustly identify the true model errors.

3. How should we test models for extreme events?

In addition to the selection bias in time, there also exists a selection bias in space. If we preselect a particular extreme event from the observational record, which, by definition, is a rare event, and look for similar events in climate model simulations, then the chances of finding rare events of the same magnitude, duration, and spatial scale, at the same location will necessarily be low. However, in this case, it is premature to then conclude that models cannot simulate the observed extreme event. For instance, if we then use a large ensemble of climate simulations and search for a similar intense event with *no a priori specification of exactly where it should happen*, it is often possible to find a similar extreme event in the simulations, leading to very different conclusions about model fidelity (e.g., Fischer et al. 2013, 2021).

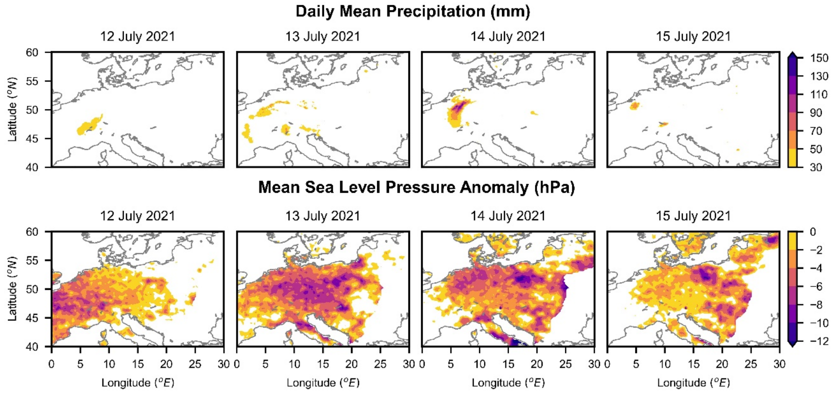

Figure 2a. Observed extreme rainfall over Europe in July 2021. Daily mean rainfall (mm) over Europe for 12-15 July (top panels) and corresponding daily mean sea level pressure anomaly (hPa) (bottom panels) from E-OBS v26.0 dataset. The daily sea level anomalies are calculated with respect to monthly values for July 2021.

To illustrate this, we use the example of the German floods of 2021. The observed event in July 2021 was associated with daily mean rainfall reaching as high as 150 mm over parts of Germany (Fig. 2a). Searching ensembles of climate simulations from multiple models for European rainfall extremes of similar magnitude to that observed, reveals several instances with rainfall intensity in climate models reaching, and in some cases even exceeding the rainfall rate seen in the observational record (Fig. 2b). We can also examine the mean sea level pressure (MSLP) to determine if the simulated extreme corresponds to a realistic circulation pattern. In all cases, the extreme rainfall events in the models are co-located with extreme low-pressure regions, indicating low-level convergence and enhanced probability of intense rainfall, similar to the observed event which had low MSLP in the vicinity.

To illustrate the point further, we also examined a large ensemble from a single model, the CESM1 (Kay et al., 2015) as this isolates the impact of internal variability. While we find similar mid-latitude extremes occurring randomly anywhere over the selected domain in some simulations of the large ensemble, other simulations from the same model did not simulate any such events. This shows that a-priori constraining the regional scale and location in the model to that of the pre-selected local record event invalidates commonly used statistical significance tests and hence can lead to the erroneous conclusion that models cannot reproduce the observed extreme.

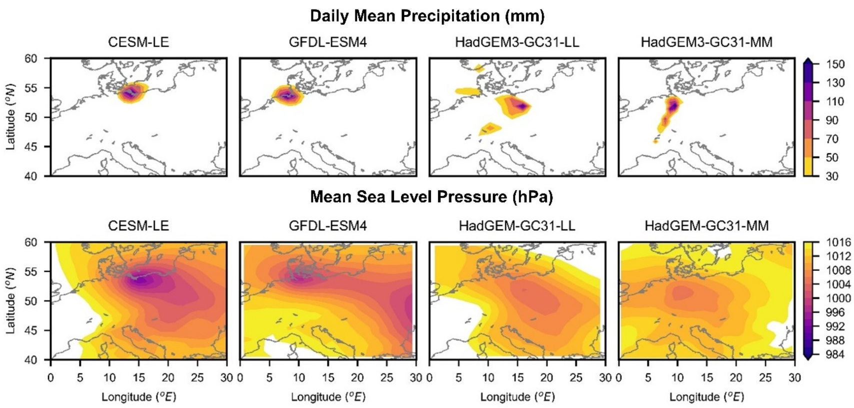

Figure 2b. Simulation of extreme European rainfall in climate models. Daily mean rainfall (mm) over Europe for a selected day between 1950-2014 (top panels). Corresponding daily mean sea level pressure (hPa) is shown in the bottom panels. The CESM-LE case is from the CESM1 Large Ensemble and the remaining models are from CMIP6 ensembles.

4. Are teleconnections changing?

Large-scale internal variability from phenomena such as ENSO, NAO, or the Indian Ocean Dipole (IOD), are known to influence regional climate across the globe through *teleconnections.* These teleconnections lead to systematic changes in regional climate far from the center of action of the variability itself (e.g., Branstator, 2002; Hurrell et al., 2003; Trenberth et al., 2020; Yamagata et al., 2004) and often contribute to the predictability of regional weather and climate events. For example, the ENSO-rainfall teleconnection is crucial for tropical rainfall prediction at seasonal lead times (Jain et al., 2019, Scaife et al., 2019).

Recent literature questions the stationarity of these and other teleconnections and in many cases argues that causal mechanisms, such as externally forced climate change, are driving systematic change in either the pattern or the strength of the teleconnection from one epoch to another, and that these changes are not represented in climate models. For instance, changes in the ENSO-rainfall teleconnection have been reported for several tropical regions including India (e.g. Krishna Kumar et al., 1999), East Asia (e.g. Wu and Wang 2002), North America (e.g. Coats et al., 2013), and Africa (e.g. Losada et al., 2012, Bahaga et al., 2019), as well as many other large-scale teleconnections, such as the recently discovered connection between the Quasi-Biennial Oscillation and Madden Julian Oscillation (Lee and Klingaman, 2018; Martin et al, 2021) and the ENSO-Asian teleconnection (He et al., 2013).

Several studies already highlight internal variability as a cause of the apparent mismatch between model and observed teleconnections (Kitoh 1997, Bodai et al., 2020, Deser et al., 2017; Lee and Bódai, 2021, Yun and Timmerman, 2018) but many others continue to suggest that mismatch implies model error. Therefore, we re-examine one well-known example: ENSO and Indian summer monsoon rainfall tele-connection.

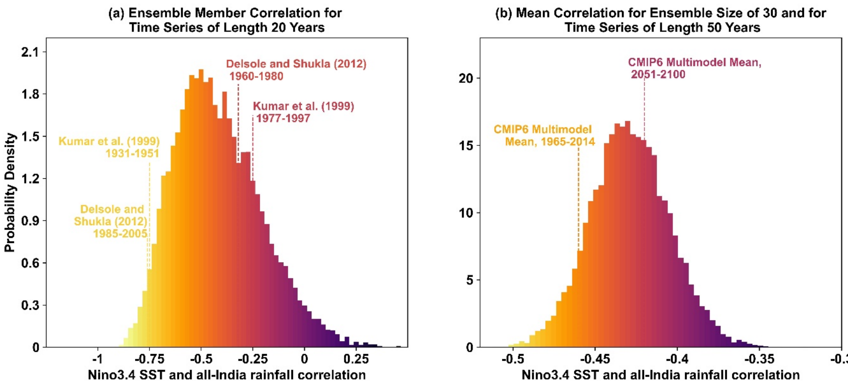

Figure 3 Internal variability in the ENSO-monsoon relationship. Probability distribution of Pearson's correlation coefficient between detrended Nino3.4 sea-surface temperatures (SSTs) and rainfall from the CHFP MME. Panel (a) is for resampled rainfall and SST time series of length 20 years and shows the range of correlations corresponding to individual ensemble member time series. Panel (b) is for 50-year time series and correlation values were randomly resampled in sets of 30 and therefore represent the range of *mean* correlations for 30 ensemble members (see Methods). The dotted lines show some extreme examples of correlation values from the literature for corresponding timescales. Examples shown in panel (b) are from Lee and Bodai (2021) for the ensemble size of 30.

Figure 3 (a) shows that the distribution of ENSO teleconnections in rainfall resamples covers an enormous range of correlations on 20-year timescales (r = -0.90 to 0.47). A similar result holds for 50-year timescales (r = -0.80 to 0.22). Note that this range occurs due to

Deleted: P

sampling variability rather than any true systematic non-stationarity, and includes the historical periods such as 1980-2010 when the non-stationarity in the ENSO-Indian rainfall teleconnection has been reported in observations (e.g. Krishna Kumar et al., 1999). Whilst methodologies to calculate the ENSO-monsoon relationship vary in the literature, even the extreme examples of apparent non-stationarity in observed teleconnections sit well within the spread of plausible ENSO-monsoon teleconnections due to the internal variability of the climate system. The model resamples even show the possibility of a positive correlation on 20-year timescales; opposite in sign to the observed teleconnection.

In addition, there is also a growing body of literature suggesting that the ENSO-Indian rainfall teleconnection could change in the future under climate change (e.g. IPCC WG1 report, Chapter 14, 2013; Roy et al., 2019). However, Fig. 3 (b) shows that the mean ENSO-rainfall correlation for the CMIP6 multimodel ensembles (of size ~30) for both historical and future periods (Lee and Bodai, 2021) sits within the range of internal variability (r = -0.31 to -0.51). Therefore, great care is needed before we can conclude with confidence that there is any robust change in the ENSO-monsoon relationship in the future.

Finally, given the broad range of possibilities in Fig. 3, and the fact that simulated future changes are well within this range, it is unlikely that observational data will yield significant examples of changes in teleconnections that are extreme enough to rule out internal variability and detect any true non-stationarity on any reasonable timescale into the future.

5. A call for more rigorous model assessment

We find that studies claiming a mismatch between model and observed climate phenomena are often too quick to ignore the null hypothesis that such apparent discrepancies between models and observations can arise due to chaotic internal variability. We have presented examples where this applies to changes on multidecadal timescales, such as global and regional trends, recent extreme events, and apparent changes in observed teleconnections.

Assessing models against the rare and most extreme observed cases automatically introduces a selection bias into the process of model evaluation which is particularly compounded for extreme events on small spatial scales. Limited observational records can easily show apparent non-stationary or spurious teleconnections due to internal variability and sampling error. Using a single simulation or small ensembles of simulations can easily lead to premature conclusions about model fidelity. Therefore, a large ensemble of multidecadal simulations is needed to properly sample initial condition uncertainty and demonstrate model failure to capture observed phenomena, or the progress or deterioration of performance between old and new generations of models, including costly new climate models at high resolution. Initialized large ensembles are already being used for seasonal predictions (Buontempo et al., 2018; Tompkins et al., 2017), and more recently for multidecadal predictions (Boer et al., 2016) and projections (Deser et al., 2020, Maher et al., 2020). Using these ensembles to isolate internal variability from true model errors provides a powerful second application.

6. What else can be done?

While a large ensemble is desirable to account for fluctuations due to internal variability, we acknowledge that these are computationally expensive and may not be always possible. Therefore, we also highlight other potential model evaluation methodologies that could be used. For instance, in contrast to picking a single period in observations and testing a single model simulation against that, we suggest using a longer time period, and sampling all possible periods of a fixed length within that interval. For example, sampling 20-year trends over 50 years for both observations and models and comparing the distribution of trends (cf. Jain and Scaife 2022, Scaife et al. 2009, Eade et al. 2022). Comparing multimodel mean trends directly with the observations is also not a fair comparison (Schmidt et al. 2023) and for these cases, statistical tests, similar to the UNprecedented Simulated Extremes in Ensembles (UNSEEN) method (Thompson et al., 2017), could be employed to test model fidelity.

Grid point comparisons for extremes, such as calculating spatial distributions of extremes (e.g., 1-day maximum rainfall) and comparing those with the model spatial distribution, are likely to show apparent disagreement. A similar problem has been recognized as 'double-penalty' in high-resolution weather prediction where verification scores are penalized twice, i.e. for simulating a feature in the wrong place and not simulating a feature at the right place (Hagelin et al., 2014). Simple approaches, such as pooling daily maximum rainfall values over a spatial domain and time period for both observations and models and comparing those distributions, or more sophisticated methodologies (such as Hi-RA, Mittermaier, 2014) could be used for evaluating climate models for extremes.

There are also new emerging methodologies, such as resampling of observational records to create pseudo observational large ensembles (McKinnon and Deser, 2018, 2021), quantification of the degree of discrepancy between models and observations using Bayes factors (Shepherd, 2021), spatial aggregation for robust projection of regional extremes using current models (Fischer et al. 2013), or using physical or process-based evaluation such as storyline methods (Shepherd et al., 2018) for post-event simulation of extreme observed events in models, that could be employed for better control on internal variability and sampling error. Finally, we highlight that models will always contain errors to some extent. However, to robustly determine true errors in climate models and to make an informed statement about them, it is important to carefully acknowledge the role of internal variability.

Methods

Figure 1 The ensemble member rainfall was bias-adjusted using the difference between the ensemble mean and observations for each model's hindcast period. Each model's fidelity was tested using the UNSEEN method and rainfall MME was created using a selected set of models that passed the UNSEEN tests (GloSea5, CFS, and the MPI models for north-central India for 1950-2013, and GloSea5, ECMWF-S4, CFS, MIROC5, and MPI-LR for all-India for 1901-2013) (see Table 1 of Jain et al, 2020 for details of models, Tompkins et al., 2017). The ensemble mean trend was removed from each ensemble member time series for each model to remove any influence of climate change before creating MME. A total of 10,000 time series of length 51 years were resampled from MME by randomly selecting ensemble members in sets of three consecutive years to preserve the interannual autocorrelation in rainfall. Linear trends were calculated as the slope of the line of best fit for each time series to obtain 10,000 values of trends. Each trend value was then multiplied by the length of the time series (i.e. 51 years) to obtain the change in JJA total rainfall over a 51-year period, shown on the x-axis. We also tested the sensitivity of the extreme values shown in Fig. 1 to the model variance by removing one of the models with the highest variance while resampling. We find no substantial influence of the differences in the models' variance on the extreme values presented in Fig. 1.

Figure 3 The rainfall MME described in Fig. 1 was used to randomly resample 10,000 time series of rainfall and corresponding sea-surface temperatures (SSTs) of lengths 20 and 50 years. The correlation coefficient was calculated for each of the 10,000 time series for both 20 (Fig. 3a) and 50 years (Figure not shown). The SSTs were averaged over the Nino3.4 region (5°S to 5°N, 170°W to 120°W) and detrended before sampling to remove the influence of forced climate change on SSTs. For Fig. 3b, we randomly selected 30 correlation values from the 10,000 correlation values for the length of 50 years. We then calculated the average of the selected 30 correlation values to obtain the *mean* correlation and this was repeated 10,000 times to obtain the distribution in Fig. 3b.

References

Bahaga, T. K., Fink, A. H., & Knippertz, P. (2019). Revisiting interannual to decadal teleconnections influencing seasonal rainfall in the Greater Horn of Africa during the 20th century. *International Journal of Climatology*, *39*(5), 2765-2785.

Bódai, T., G. Drótos, M. Herein, F. Lunkeit, and V. Lucarini (2020), The Forced Response of the El Niño–Southern Oscillation–Indian Monsoon Teleconnection in Ensembles of Earth System Models, Journal of Climate, 33(6), 2163-2182.

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., ... & Eade, R. (2016). The decadal climate prediction project (DCPP) contribution to CMIP6. Geoscientific Model Development, 9(10), 3751-3777.

Branstator, G. (2002). Circumglobal Teleconnections, the Jet Stream Waveguide, and the North Atlantic Oscillation, Journal of Climate, 15(14), 1893-1910.

Broecker, W. S. (1975). Climatic change: are we on the brink of a pronounced global warming?. Science, 189(4201), 460-463.

Buontempo, C., Hewitt, C. D., Doblas-Reyes, F. J., & Dessai, S. (2014). Climate service development, delivery and use in Europe at monthly to inter-annual timescales. Climate Risk Management, 6, 1-5.

Butler, A. H., & Polvani, L. M. (2011). El Niño, La Niña, and stratospheric sudden warmings: A reevaluation in light of the observational record. Geophysical Research Letters, 38(13).

Coats, S., Smerdon, J. E., Cook, B. I., & Seager, R. (2013). Stationarity of the tropical pacific teleconnection to North America in CMIP5/PMIP3 model simulations. Geophysical Research Letters, 40(18), 4927-4932.

DelSole, T., & Shukla, J. (2012). Climate models produce skillful predictions of Indian summer monsoon rainfall. Geophysical Research Letters, 39(9).

Deser, C. (2020). “Certain uncertainty: The role of internal climate variability in projections of regional climate change and risk management”. Earth's Future, 8, e2020EF001854. https://doi.org/10.1029/2020EF001854

Deser, C. and A. S. Phillips, 2023: A range of outcomes: The combined effects of internal variability and anthropogenic forcing on regional climate trends over Europe. *EGU Nonlinear Proc. Geoph. Special Issue* (accepted and in press as a Highlight Article).

Deser, C., I. R. Simpson, K. A. McKinnon and A. S. Phillips (2017). The Northern Hemisphere extra-tropical atmospheric circulation response to ENSO: How well do we know it and how do we evaluate models accordingly? J. Climate, 30, 5059-5082, doi:10.1175/JCLI-D-16-0844.1.

Deser, C., L. Terray and A. S. Phillips (2016). Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications. J. Climate, 29, 2237-2258, doi:10.1175/JCLI-D-15-0304.1.

Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N. et al. (2020). Insights from Earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, *10*(4), 277-286.

Deser, C., R. Knutti, S. Solomon, and A. S. Phillips (2012). Communication of the role of natural variability in future North American climate. Nat. Climate Change, 2, 775–779, https://doi.org/10.1038/nclimate1562.

Deser, C., Simpson, I. R., Phillips, A. S., & McKinnon, K. A. (2018). How well do we know ENSO's climate impacts over North America, and how do we evaluate models accordingly?. Journal of Climate, 31(13), 4991-5014.

Eade, R., Stephenson, D. B., Scaife, A. A., & Smith, D. M. (2022). Quantifying the rarity of extreme multi-decadal trends: how unusual was the late twentieth century trend in the North Atlantic Oscillation?. Climate Dynamics, 58(5), 1555-1568.

Fasullo, J., A. S. Phillips and C. Deser, (2020). Evaluation of Leading Modes of Climate Variability in the CMIP Archives. J. Climate, 33, 5527–5545, doi:10.1175/JCLI-D-19-1024.1.

Fischer, E. M., Sippel, S., & Knutti, R. (2021). Increasing probability of record-shattering climate extremes. Nature Climate Change, 11(8), 689-695.

Fischer, E., Beyerle, U. & Knutti, R. Robust spatially aggregated projections of climate extremes. Nature Clim Change 3, 1033–1038 (2013). https://doi.org/10.1038/nclimate2051

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S.C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason and M. Rummukainen (2013). Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Gillett, N. P., Zwiers, F. W., Weaver, A. J., & Stott, P. A. (2003). Detection of human influence on sea-level pressure. *Nature*, *422*(6929), 292-294.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N., & Tennant, W. (2017). The Met Office convective-scale ensemble, MOGREPS-UK. Quarterly Journal of the Royal Meteorological Society, 143(708), 2846-2861.

Hansen, J., Johnson, D., Lacis, A., Lebedeff, S., Lee, P., Rind, D., & Russell, G. (1981). Climate impact of increasing atmospheric carbon dioxide. Science, 213(4511), 957-966.

Hardiman, S. C., Dunstone, N. J., Scaife, A. A., Smith, D. M., Knight, J. R., Davies, P., ... & Greatbatch, R. J. (2020). Predictability of European winter 2019/20: Indian Ocean dipole impacts on the NAO. *Atmospheric Science Letters*, *21*(12), e1005.

Hausfather, Z., Drake, H. F., Abbott, T., & Schmidt, G. A. (2020). Evaluating the performance of past climate model projections. Geophysical Research Letters, 47(1), e2019GL085378.

He, S., Wang, H., & Liu, J. (2013). Changes in the relationship between ENSO and Asia–Pacific midlatitude winter atmospheric circulation. *Journal of Climate*, *26*(10), 3377-3393.

Huang, X., Zhou, T., Turner, A., Dai, A., Chen, X., Clark, R., ... & Zou, L. (2020). The recent decline and recovery of Indian summer monsoon rainfall: relative roles of external forcing and internal variability. Journal of Climate, 33(12), 5035-5060.

Hurrell, J.W., Kushnir, Y., Ottersen, G. and Visbeck, M. (2003). An Overview of the North Atlantic Oscillation. In The North Atlantic Oscillation: Climatic Significance and Environmental Impact (eds J.W. Hurrell, Y. Kushnir, G. Ottersen and M. Visbeck). https://doi.org/10.1029/134GM01

IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., 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, and B. Zhou (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2391 pp. doi:10.1017/9781009157896.

Jain S, Scaife AA and Mitra AK (2019). Skill of Indian summer monsoon rainfall prediction in multiple seasonal prediction systems Climate Dynamics 52(9-10) 5291-5301.

Jain, S., & Scaife, A. A. (2022). How extreme could the near term evolution of the Indian Summer Monsoon rainfall be?. Environmental Research Letters, 17(3), 034009.

Jain, S., Scaife, A. A., Dunstone, N., Smith, D., & Mishra, S. K. (2020). Current chance of unprecedented monsoon rainfall over India using dynamical ensemble simulations. Environmental Research Letters, 15(9), 094095.

Jin, Q. and Wang, C. (2017). A revival of Indian summer monsoon rainfall since 2002. Nature Climate Change, 7(8), pp.587-594.

Katzenberger, A., Schewe, J., Pongratz, J., & Levermann, A. (2021). Robust increase of Indian monsoon rainfall and its variability under future warming in CMIP6 models. Earth System Dynamics, 12(2), 367-386.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J. F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., & Vertenstein, M. (2015). The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. Bulletin of the American Meteorological Society, 96(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Kent, C., Dunstone, N., Tucker, S., Scaife, A. A., Brown, S., Kendon, E. J. et al. (2022). Estimating unprecedented extremes in UK summer daily rainfall. *Environmental Research Letters*, *17*(1), 014041.

Kitoh, A. (2007). Variability of Indian monsoon-ENSO relationship in a 1000-year MRI-CGCM2. 2 simulation. *Natural Hazards*, *42*(2), 261-272.

Kumar, K. K., Rajagopalan, B., & Cane, M. A. (1999). On the weakening relationship between the Indian monsoon and ENSO. *Science*, *284*(5423), 2156-2159.

Kumar, V., Jain, S. K., & Singh, Y. (2010). Analysis of long-term rainfall trends in India. Hydrological Sciences Journal–Journal des Sciences Hydrologiques, 55(4), 484-496.

Lee, J. C., & Klingaman, N. P. (2018). The effect of the quasi-biennial oscillation on the Madden–Julian oscillation in the Met Office Unified Model Global Ocean Mixed Layer configuration. *Atmospheric Science Letters*, *19*(5), e816.

Lee, J. Y., & Bódai, T. (2021). Future changes of the ENSO–Indian summer monsoon teleconnection. In Indian Summer Monsoon Variability (pp. 393-412). Elsevier.

LeGrande, A. N., Tsigaridis, K., & Bauer, S. E. (2016). Role of atmospheric chemistry in the climate impacts of stratospheric volcanic injections. Nature Geoscience, 9(9), 652-655.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. Journal of the Atmospheric Sciences, 20(2), 130– 141. https://doi.org/10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2

Lorenz, E. (1969). The predictability of a flow which posses many scales of motion. Tellus, 21(3), 289– 307. https://doi.org/10.1111/j.2153-3490.1969.tb00444.x

Losada, T., Rodriguez-Fonseca, B., Mohino, E., Bader, J., Janicot, S., and Mechoso, C. R. (2012), Tropical SST and Sahel rainfall: A non-stationary relationship, Geophys. Res. Lett., 39, L12705.

Maher, N., Lehner, F., & Marotzke, J. (2020). Quantifying the role of internal variability in the temperature we expect to observe in the coming decades. Environmental Research Letters, 15(5), 054014.

Martin, Z., Orbe, C., Wang, S., & Sobel, A. (2021). The MJO–QBO Relationship in a GCM with Stratospheric Nudging. Journal of Climate, 34(11), 4603-4624.

McKinnon, K. A. and C. Deser (2021), The inherent uncertainty of precipitation variability, trends, and extremes due to internal variability, with implications for Western US water resources. J. Climate, 34, 9605-9622, doi: 10.1175/JCLI-D-21-0251.1.

McKinnon, K. A., & Simpson, I. R. (2022). How unexpected was the 2021 Pacific Northwest heatwave?. Geophysical Research Letters, 49(18), e2022GL100380.

Mishra, S. K., Sahany, S., Salunke, P., Kang, I. S., & Jain, S. (2018). Fidelity of CMIP5 multi-model mean in assessing Indian monsoon simulations. npj Climate and Atmospheric Science, 1(1), 1-8.

Mitchell, D. M., Thorne, P. W., Stott, P. A., & Gray, L. J. (2013). Revisiting the controversial issue of tropical tropospheric temperature trends. Geophysical Research Letters, 40(11), 2801-2806.

Mitchell, J. F. B., Johns, T. J., Gregory, J. M., & Tett, S. F. B. (1995). Transient climate response to increasing sulphate aerosols and greenhouse gases. *Nature*, *376*, 501-504.

Mittermaier, M. P. (2014). A strategy for verifying near-convection-resolving model forecasts at observing sites. Weather and Forecasting, 29(2), 185-204.

Monerie, P. A., Wilcox, L. J., & Turner, A. G. (2022). Effects of anthropogenic aerosol and greenhouse gas emissions on Northern Hemisphere monsoon precipitation: mechanisms and uncertainty. Journal of Climate, 1-66.

Paul, S., Ghosh, S., Oglesby, R., Pathak, A., Chandrasekharan, A., & Ramsankaran, R. A. A. J. (2016). Weakening of Indian summer monsoon rainfall due to changes in land use land cover. *Scientific reports*, *6*(1), 1-10.

Rahmstorf, S., Foster, G., & Cahill, N. (2017). Global temperature evolution: recent trends and some pitfalls. *Environmental Research Letters*, *12*(5), 054001.

Ramesh, K. V., & Goswami, P. (2014). Assessing reliability of regional climate projections: the case of Indian monsoon. *Scientific reports*, *4*(1), 1-9.

Roxy, M. K., Ritika, K., Terray, P., Murtugudde, R., Ashok, K., & Goswami, B. N. (2015). Drying of Indian subcontinent by rapid Indian Ocean warming and a weakening land-sea thermal gradient. Nature communications, 6(1), 1-10.

Saha, A., S. Ghosh, A. S. Sahana, and E. P. Rao (2014), Failure of CMIP5 climate models in simulating post-1950 decreasing trend of Indian monsoon, Geophysical Research Letters, 41(20), 7323-7330.

Santer, B. D., Thorne, P. W., Haimberger, L., Taylor, K. E., Wigley, T. M. L., Lanzante, J. R., ... & Wentz, F. J. (2008). Consistency of modelled and observed temperature trends in the tropical troposphere. International Journal of Climatology: A Journal of the Royal Meteorological Society, 28(13), 1703-1722.

Scafetta, N. (2022). Advanced Testing of Low, Medium, and High ECS CMIP6 GCM Simulations Versus ERA5-T2m. Geophysical Research Letters, 49(6), e2022GL097716.

Scaife, A. A., Ferranti, L., Alves, O., Athanasiadis, P., Baehr, J., Dequé, M. et al. (2019). Tropical rainfall predictions from multiple seasonal forecast systems. *International Journal of Climatology*, *39*(2), 974-988.

Schmidt, G. A., G. S. Jones and J. J. Kennedy, 2023. Comment on “Advanced Testing of Low, Medium, and High ECS CMIP6 GCM Simulations Versus ERA5-T2m” by N. Scafetta (2023), Geophys. Res. Lett., doi:10.1029/2022GL102530 (in press).

Schmidt, G. A., Shindell, D. T., & Tsigaridis, K. (2014). Reconciling warming trends. Nature Geoscience, 7(3), 158-160.

Seidel, D. J., Fu, Q., Randel, W. J., & Reichler, T. J. (2008). Widening of the tropical belt in a changing climate. *Nature geoscience*, *1*(1), 21-24.

Shashikanth, K., Salvi, K., Ghosh, S. and Rajendran, K. (2014), Do CMIP5 simulations of Indian summer monsoon rainfall differ from those of CMIP3?. Atmos. Sci. Lett., 15: 79-85.

Shepherd, T.G. (2021). Bringing physical reasoning into statistical practice in climate-change science. Climatic Change, 169, 2.

Shepherd, T.G., Boyd, E., Calel, R.A., Chapman, S.C., Dessai, S., Dima-West, I.M., Fowler, H.J., James, R., Maraun, D., Martius, O., Senior, C.A., Sobel, A.H., Stainforth, D.A., Tett, S.F.B., Trenberth, K.van den Hurk, B.J.J.M., Watkins, N.Wilby, R.L. and Zenghelis, D.A. (2018). Storylines: an alternative approach to representing uncertainty in physical aspects of climate change. Climatic Change, 151, 555–571.

Singh, D., Ghosh, S., Roxy, M. K., & McDermid, S. (2019). Indian summer monsoon: Extreme events, historical changes, and role of anthropogenic forcings. Wiley Interdisciplinary Reviews: Climate Change, 10(2), e571.

Sinha, A., Kathayat, G., Cheng, H., Breitenbach, S. F., Berkelhammer, M., Mudelsee, M. et al. (2015). Trends and oscillations in the Indian summer monsoon rainfall over the last two millennia. *Nature communications*, *6*(1), 1-8.

Slingo, J., Bates, P., Bauer, P., Belcher, S., Palmer, T., Stephens, G., et al. (2022). Ambitious partnership needed for reliable climate prediction. Nature Climate Change, 12(6), 499-503.

Swart, N. C., Fyfe, J. C., Hawkins, E., Kay, J. E., & Jahn, A. (2015). Influence of internal variability on Arctic sea-ice trends. Nature Climate Change, 5(2), 86-89.

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. Philosophical transactions of the royal society A: mathematical, physical and engineering sciences, 365(1857), 2053-2075.

Thompson V, Dunstone NJ, Scaife AA, Smith DM, Slingo JM, Brown S and Belcher SE (2017) High risk of unprecedented UK rainfall in the current climate Nature communications, 8(1), 107.

Tompkins, A. M., De Zárate, M. I. O., Saurral, R. I., Vera, C., Saulo, C., Merryfield, W. J., et al. (2017). The climate-system historical forecast project: providing open access to seasonal forecast ensembles from centers around the globe. Bulletin of the American Meteorological Society, 98(11), 2293-2301.

Trenberth, K.E. (2020). ENSO in the Global Climate System. In El Niño Southern Oscillation in a Changing Climate (eds M.J. McPhaden, A. Santoso and W. Cai). https://doi.org/10.1002

Tyrrell, N. L., Koskentausta, J. M., & Karpechko, A. Y. (2022). Sudden stratospheric warmings during El Niño and La Niña: sensitivity to atmospheric model biases. Weather and Climate Dynamics, 3(1), 45-58.

Wu, R., & Wang, B. (2002). A Contrast of the East Asian Summer Monsoon–ENSO Relationship between 1962–77 and 1978–93, Journal of Climate, 15(22), 3266-3279.

Yamagata, T., Behera, S. K., Luo, J. J., Masson, S., Jury, M. R., & Rao, S. A. (2004). Coupled ocean-atmosphere variability in the tropical Indian Ocean. Earth's Climate: The Ocean–Atmosphere Interaction, Geophys. Monogr, 147, 189-212.Yun, K. S., & Timmermann, A. (2018). Decadal monsoon-ENSO relationships reexamined. Geophysical Research Letters, 45(4), 2014-2021.

Acknowledgments

SJ and AAS jointly wrote the first draft. All co-authors have contributed to the editing and writing of the manuscript. The data that support the findings of this study are openly available at the following URLs: https://cds.climate.copernicus.eu/cdsapp#!/dataset/projections-cmip6, https://www.wcrp-climate.org/wgsip-chfp/chfp-data-archive and https://www.cesm.ucar.edu/projects/community-projects/LENS/data-sets.html. The IMD rainfall data can be obtained from the following link: https://www.imdpune.gov.in/Clim_Pred_LRF_New/Grided_Data_Download.html.

The authors declare no conflict of interest.