Does statistical model perform at par with computationally expensive general circulation model for decadal prediction?

Rishi Sahastrabuddhe and Subimal Ghosh

1 Department of Civil Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India
2 Interdisciplinary Program in Climate Studies, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India

* Author to whom any correspondence should be addressed.

E-mail: subimal@civil.iitb.ac.in

Keywords: decadal predictions, limits of predictability, computationally inexpensive data-driven models, Indian summer monsoon rainfall

Abstract

Decadal predictions have gained immense importance over the last decade because of their use in near-term adaption planning. Computationally expensive coupled model intercomparison project phase 5 general circulation models (GCMs) are initialized every 5 years and they generate the decadal hindcasts with moderate skill. Here we test the hypothesis that computationally inexpensive data-driven models, such as multi-variate singular spectrum analysis (MSSA), which takes care of trends and oscillations, performs similar to GCMs. We pick up one of the most complex variables having low predictability, Indian summer monsoon rainfall (ISMR) and its possible causal sea surface temperatures (SST). We find that the MSSA approach performs similar to the GCMs in simulating SSTs beyond their nonlinear limits of predictability, which is ∼12 months. These SSTs are used for decadal predictions of ISMR and show improved skills compared to the GCMs. We conclude that data-driven models are possible alternatives to computationally expensive GCMs for decadal predictions.

1. Introduction

Decadal prediction has gained attention in recent times for its usefulness in short-term adaptation and mitigation planning. This provides enough lead time to the policymakers and stakeholders to plan for short term adaptation strategies. Coupled general circulation models (GCMs) attempt to solve the fascinating challenge of decadal prediction simulating a variety of complex mechanisms and modeling physical processes from differential equations (DelSole and Tippett 2007, Smith et al 2007, Keenlyside et al 2008, Pohlmann et al 2009, van Oldenborgh et al 2012). For the decadal predictions, the GCMs are initialized with the climate states of a specific year and are run for the next 10 years. This is as per the experimental design specified by the coupled model intercomparison project 5 (CMIP5) (Taylor et al 2012). Though they are computationally expensive, at present, they are the only available tools for decadal predictions. However, the decadal prediction skills remain low to moderate to be used for adaptation by the broad groups of stakeholders (Knutti and Sendlácek 2013, Hidalgo and Alfaro 2015). Existing literature provides a little understanding towards the arbitrary skills for multi-year/decadal predictions of GCMs (Branstator et al 2012, Delsole et al 2013, Buckley et al 2019).

In a decadal prediction system, GCMs attempt to trace the evolving decadal climate from the initial conditions and initial state. The expected predictive skills of these models are associated with the underlying oceanic low-frequency variability (Hawkins and Sutton 2009). However, there exists the lack of capability in these models to simulate climate variables (Lienert and Doblas-Reyes 2013b). Errors in initial conditions and climate drift inevitably manifest and corrupt any potential maximization of long-term predictive skills from the models (Lorenz 1963b, DelSole and Tippett 2008). Among the predicted variables, the low-frequency variables such as sea surface temperatures (SSTs) are expected to show higher skills (Kim et al 2011, van Oldenborgh et al 2012, Meehl...
This is because their predictability resides in the slowly evolving subsurface oceanic component and events such as El Niño (Chikamoto et al 2013). However, the predictions of SSTs, such as El Niño Southern Oscillation (ENSO), are only usable for a limited lead time of 6–12 months (Collins et al 2002, Barnston et al 2012). The low-frequency subsurface variability propagates into the atmosphere and land through ocean–atmosphere coupling, modulating surface temperatures, precipitation, surface runoffs and other processes (Enfield et al 2001, Sutton 2005, Knight et al 2006). These predictive skills drop drastically over the land (Meehl et al 2014) due to the high-frequency variability of land variables with short-term memory. Hence, in a coupled ocean–atmosphere decadal prediction model, the predictability of land and atmospheric variables remains low, though the models themselves are computationally very demanding.

Data-driven statistical models have emerged as an alternative to complex and computationally expensive dynamical climate models. An ideal statistical model should be exclusively data-driven while capturing the characteristics of a climate model, but the quality and length of the data have remained significant limitations for their development (Ho et al 2013). Recent literature also proposed the idea of physics guided data-driven models (Ganguly et al 2014, Jia et al 2020, Kodra et al 2020), where the data-driven relationship may also be partially guided and constrained by physics, which adds strength to a simple statistical model. The data-driven models are based on scientific knowledge that is often inadequate with number of caveats (Jia et al 2020). The objective of physics guided data-driven modelling technique is to bring the power of both physics models and data-driven methods, where the physics guided system conceptually represents lower dimensional representations, while data models (especially recent incarnations such as deep learning) represent higher dimensional representations. Recent AI/ML methods show promising results in climate applications (Scher 2018, Huntingford et al 2019, Reichstein et al 2019). The guidance of physics may come at different steps such as covariate selections with physics and statistics (Constantinescu and Anitescu 2013), modelling of the residuals (Karpatne et al 2017b), bias correction (Jia et al 2020), partial or full replacement of specific processes (Willard et al 2020), physical constraints on cost functions or on the outputs (Karpatne et al 2017a, Elhamod et al 2020), and nudging towards approximate solutions (Pawar et al 2020), etc. Despite the recent advancements in data-driven modelling mentioned above, coupled climate models outperforming the statistical models remains the current status of climate simulations or prediction system (McPhaden et al 2006). Several studies attempted to predict the long-term slowly evolving climate states using data-driven models (Lean and Rind 2009b, Krueger and Von Storch 2011, Ho et al 2013). The changes in climate on decadal scale are largely attributed to the radiative forcing and its internal variability (Keenlyside and Ba 2010, Solomon et al 2011, Goddard et al 2012b). GCMs are able to show good prediction skills for non-stationary climate with long-term forced trend (Lee et al 2006b, van Oldenborgh et al 2012). The skills of prediction from the internal variability are worth exploring. Ho et al (2013) used the global mean equivalent CO₂ concentration to represent the trend component from SSTs in a regression model. The residuals of the model represent the internal variability in SSTs. They identified different regions where the forced trend and internal variability play a significant role in skillful decadal predictions. But the decadal prediction of complex atmospheric processes, like monsoon processes, has remained unexplored. These processes generally have a short-term memory, and hence, the models generate no skill for the decadal prediction. Potential predictability of precipitation can be exploited using a statistical model from the modes of decadal variability and consequent feedback of SST (Sun et al 2015, Redolat et al 2019).

Indian summer monsoon rainfall (ISMR) is a complex response to the time-varying interactions from intra-seasonal regional-scale events (Webster et al 1998, Gadgil 2003) to interannually varying factors (Kumar et al 1995) and multi-decadal large-scale forcings (Krishnan and Sugi 2003, Zhang and Delworth 2006, Krishnamurthy and Krishnamurthy 2014a, 2014b, 2016). Around 80% population of India depends on ISMR for agricultural yield and socioeconomic activities. This requires skillful decadal-scale predictability of ISMR for planning developmental activities. The multi-year/decadal predictability of Indian monsoon is more likely linked to the slow evolution of SST variability (Suckling and Smith 2013). In CMIP5 models, SSTs modulate the Indian monsoon mechanism through the atmosphere–ocean coupling (Lu et al 2006, Malik et al 2017). The interannual projections of ISMR have shown large inter-model spread and consistent increase under persistent global warming (Fan et al 2012, Menon et al 2013), with the drawbacks in capturing the spatial heterogeneity accurately (Turner and Annamalai 2012). Very few CMIP5 models based on ensemble means are found simulating the monsoon intra-seasonal characteristics well. The improvement in the convective parameterization scheme of CMIP5 models can help capture the monsoon characteristics (Sabeerali et al 2013). There are biases in the ISMR magnitude, though the ensemble means overall capture characteristics of observed ISMR (Mishra et al 2018). The greenhouse gases (GHG)-induced warming and the direct and indirect impacts of anthropogenic aerosols on the recent trend of ISMR have been studied extensively in the literature (Bollasina et al 2011, Guo et al 2015, Rahastrabuddhe and S Ghosh)
Roxy et al., 2015). CMIP5 models reproduce a better representation of observed land-precipitation trend under natural and anthropogenic aerosol forcings, along with its inter-decadal variability under indirect effect (Wilcox et al., 2013). However, the coupled models have failed to address the observed long-term characteristics of ISMR, such as, decreasing trend of ISMR (Turner and Annamalai, 2012, Saha et al., 2014). They produce dry bias over Central India (Sperber et al., 2013, Sahana et al., 2019), lacking skill of land–atmosphere interactions (Devanand et al., 2018) and opposite trends due to the missing representation of aerosols (Bollasina et al., 2011); although, the impacts of absorbing aerosol over ISMR is debated by Lau and Kim (2011). Moreover, the models are unable to capture the newer predictors of ISMR in a warming climate resulting in a poor predictability (Wang et al., 2015). These model characteristics cause unreliable decadal predictions of ISMR with the huge inter-model spread and faulty skillset (Annamalai et al., 2007, Kripalani et al., 2007, Turner and Slingo, 2009, Saha et al., 2014). Hence, ISMR decadal predictions from coupled models remained more experimental rather than being operational.

We believe that the lack of capability of dynamical or statistical models for long-term predictions necessitates an improved understanding of limits to predictability. The predictive skill of any model, in simulating any geophysical variable, can only hold good up to a specific period (limit of predictability). Beyond the limit of predictability, the physical or statistical models may show similar predictive skills. Here, we test our hypothesis with the unique component of the global monsoon system, i.e. the ISMR and its associated SST drivers. A recent analysis (Sahastrabuddhe et al., 2019) showed that the interannual variations and the spatial patterns of ISMR are associated with fluctuations of SSTs over 12 oceanic regions. In the present analysis, we consider the same 12 regions and they are listed in figure 1. The

Figure 1. (a) Twelve oceanic regions considered in the study. (b) Limits of predictability of SSTs over the 12 oceanic regions (boxes in (a) excluding land areas).
Table 1. CMIP5 GCMs used for decadal prediction in the present study.

| Global climate model (CMIP5) | Resolution | Hindcast ensemble size |
|------------------------------|------------|------------------------|
| Beijing Climate Center, Climate System Model, version 1.1 (BCC-CSM1.1) | 2.8° | 1° × 0.776° | 4 |
| Fourth Generation Canadian Coupled Atmosphere Ocean General Circulation Model (CanCM4) | 2.8° | 1° × 0.9375° | 10 |
| Centre National de Recherches Meteorologiques Coupled Global Climate Model, version 5 (CNRM-CM5) | 1.4° | 1° × 0.6164° | 10 |
| Hadley Centre Coupled Model, version 3 (HadCM3) | 2.5° × 3.75° | 1.25° | 10 |
| L’Institut Pierre-Simon Laplace Coupled Model, version 5, coupled with NEMO, low resolution (IPSL-CM5A-LR) | 1.9° × 3.75° | 1.25° | 6 |
| Model for Interdisciplinary Research on Climate, version 5 (MIROC5) | 1.4° | 1° × 0.8° | 6 |
| Max Planck Institute Earth System Model, low resolution (MPI-ESM LR) | 1.9° | 1° × 0.8° | 10 |
| Meteorological Research Institute Coupled Atmosphere Ocean General Circulation Model, version 3 (MRI-CGCM3) | 1.1° | 1° × 0.5° | 9 |

oceanic regions are—Atlantic Niño (AtlNiño), Bay of Bengal (BoB), Indian Ocean Dipole (IOD), Niño3, Niño3.4, Niño4, North Atlantic (NA), North Pacific 1 (NP1), North Pacific 2 (NP2), South China Sea (SCS), Southern Indian Ocean (SIO), and Western Indian Ocean (WestIO). We first quantify the nonlinear limits of predictability for the SSTs over these 12 regions using nonlinear local Lyapunov exponent (NLLE). We explore the underlying variability of SST using a signal analysis technique, multivariate singular spectrum analysis (MSSA) (Golyandina et al 2001). The technique was previously used by Krishnamurthy and Shukla (2008) to understand the spatio-temporal structure of the intraseasonal variability of Indian monsoon rainfall. Using MSSA, they found that the 45 and 28 d of intraseasonal oscillatory modes can explain the active and break cycles of the Indian monsoon. Decadal predictions for SSTs are made using MSSA and the skills are compared with those obtained from eight GCMs of CMIP5 suite. The GCMs are selected based on output availability and are listed in table 1. We further predict the ISMR with two data-driven models. In the first model, we have applied singular spectrum analysis (SSA) directly to the ISMR (as we have only one variable to perform the predictions) for its decadal prediction. This is purely a data-driven approach. In the second model, the ISMR is predicted at a decadal scale with the inputs as the predicted SSTs from the MSSA model. The selection of oceanic regions for the SST predictors are guided by our physical understanding of monsoon and, hence, we may consider the second model as physics guided data-driven model. We compare the prediction skills from both the methods to those obtained from GCMs used for decadal prediction. The comparative skills of the two data-driven models further show that the guidance of physics (in terms of identifying SST regions and using them) improves the performance of a statistical model purely driven by data.

2. Data and methods

2.1. Rainfall and sea surface temperature (SST) data

The monthly statistical means from the Extended Reconstructed SST version 3b (ERSST v3b) is used as the observed SST data. Comprehensive ocean–atmosphere data set of SST is obtained for the period 1901–2015 at a resolution of 2° × 2° (Smith et al 2008).

The seasonal mean rainfall of ISMR for the 122 d in a year, viz., 1 June–30 September, is used in the analysis. We use high spatial resolution (0.25° × 0.25°) daily gridded dataset over India from India Meteorological Department (IMD) for the long-term of 1901–2015 (Pai et al 2014).

Monthly decadal hindcasts/forecasts are obtained from eight CMIP5 GCMs, as given in table 1. We choose these eight models as per the availability of the decadal outputs for both SST and precipitation variables. The decadal predictions initialized at the end of 1960, 1965, 1970, …, 2005 are extended for the next 10 years. The initializations are expected to start no earlier than 4 months before the years of interest. Observed anomalies or full field initializations are required for oceans in CMIP5 GCMs. However, land, sea-ice, and atmosphere initializations are left to each group’s decision to run decadal predictions, given that the observed state of the climate is
represented on the start date. The decadal predictions include observed ‘All forcings’ with concentrations of well-mixed GHGs, ozone, and aerosols (along with aerosols from volcanic eruptions) (Taylor et al 2012). We extract 5 year interval hindcasts initialized in 1960, 1965, 1970, .., 2005. Statistical mean of all the simulations by a model for a specific initialization represents that particular model in evaluating the performance. We have used the available ensemble hindcast members (table 1) and computed the ensemble mean. The ensemble mean is used for the analysis. The forecasted realizations start from January of the starting of a decade to the December of the end of the same decade for all the models, except HadCM3, for which the forecasts are from November of the initialized year. For each coupled model, the monthly means of SST outputs are re-gridded to the resolution of the observed ERSST v3b dataset (i.e. 2° × 2°) using bilinear interpolation. Monthly means of precipitation from the CMIP5 models have been analyzed over India for June–September without any re-gridding. As we are considering spatial average of ISMR, re-gridding is not needed.

Monthly SST data from ERSST v3b and ISMR from IMD serve as the observed datasets to compare skills and to develop the data-driven model.

2.2. Nonlinear local lyapunov exponent

NLLE, an improved version of classical Lyapunov exponent, is used to determine both linear and nonlinear initial error growth rates of a system. Initially developed by Chen et al (2006), the method was further mathematically established by Ding and Li (2007). Herewith, we discuss the general form of the technique along with the drawbacks of the older tangent linear model for approximation of error growth rates.

Consider an \(m\)-dimensional continuous-time dynamical system:

\[
\frac{dY(t)}{dt} = G(Y(t))
\]

where \(Y = [y_1(t), y_2(t), y_3(t), \ldots, y_m(t)]^T\) is the state vector at the time \(t\), the superscript \(T\) is the transpose, and \(G\) represents the dynamics of a system. The evolution of a small error \(\varepsilon = [\varepsilon_1(t), \varepsilon_2(t), \varepsilon_3(t), \ldots, \varepsilon_m(t)]^T\), superimposed on a state \(y\), is governed by the nonlinear equations:

\[
\frac{d\varepsilon}{dt} = A(Y)\varepsilon + B(Y, \varepsilon)
\]

where \(A(Y)\varepsilon\) depicts the tangent linear terms and \(B(Y, \varepsilon)\) are the high-order nonlinear terms of the error \(\varepsilon\). Earlier studies sought to approximate the evolution of error growth by the tangent linear model without considering the nonlinear part of the equation, assuming initial errors were sufficiently small (Lorenz 1965, Ziehmann et al 2000). However, the ignored nonlinear errors grow exponentially, adding to the evolution of the total error growth rate (Chen et al 2006, Ding and Li 2007). Moreover, the linear approximation does not apply to the dynamics with a considerable initial error. Hence, the consideration of nonlinear error growth rate is needed to determine the overall limit of predictability of a system.

Without a linear approximation, the solutions of equation (2) can be obtained by numerically integrating it along with the reference solution of \(y\) from \(t = t_0\) to \(t_0 + \tau\) (\(\tau\) is a small interval):

\[
\varepsilon_1 = \eta(y_0, \varepsilon_0, \tau)\varepsilon_0
\]

where \(\varepsilon_1 = (t_0 + \tau)\) and \(\eta(y_0, \varepsilon_0, \tau)\) is a nonlinear propagator.

The NLLE is defined as:

\[
\lambda(y_0, \varepsilon_0, \tau) = \frac{1}{\tau} \ln \left( \frac{\|\varepsilon_1\|}{\|\varepsilon_0\|} \right)
\]

where \(\eta(y_0, \varepsilon_0, \tau)\varepsilon_0\) depends on the initial state \(y_0\), the initial error \(\varepsilon_0\) and time \(\tau\). NLLE differs from classical Lyapunov exponents, which is defined based on tangent linear error dynamics and is derived solely from the initial state \(y_0\) and time \(\tau\), without considering the initial error \(\varepsilon_0\). The mean NLLE of the dynamical system is:

\[
\bar{\lambda}(\varepsilon_0, \tau) = \int_{\Omega} \lambda(y_0, \varepsilon_0, \tau) \, dx
\]

where \(\Omega\) is the domain for the global attractor of the system dynamics with a sufficiently large sample size. The mean relative growth of initial error can be denoted as:

\[
\hat{\Phi}(\varepsilon_0, \tau) = \exp \left[ \bar{\lambda}(\varepsilon_0, \tau) \tau \right].
\]

From the theorem in Ding and Li (2007),

\[
\hat{\Phi}(\varepsilon_0, \tau) \to c(N \to \infty).
\]

The constant \(c\) is the theoretical saturation level that depends on the converged probability of the error growth of the system. At the saturation level, all the information from the initial state of the system is lost. The prediction system developed with the initial state becomes meaningless after that. We try to find these nonlinear limits of predictability of SSTs over 12 oceanic regions, which are the excellent predictors for ISMR, as mentioned in Sahastrabuddhe et al (2019). Monthly observed SST data over the period of 1901–2010, viz., time series of 1320 values over each SST region, are analyzed to demonstrate the limits using NLLE. The limits over the regions are measured at a 95% saturation level of error growth. A step by step algorithm for the estimation of NLLE is discussed in supplementary text S1 (available online at stacks.iop.org/ERL/16/064028/mmedia).
2.3. Evaluation of decadal prediction skills
The decadal predictive skill of simulated SST by the climate models is evaluated with the metric, anomaly correlation coefficient (ACC), where the simulated SSTs are compared to the observed SSTs. The equation for computing ACC is as follows:
\[
\text{ACC} = \frac{\sum_{j=1}^{n} [Y_j - \bar{Y}] [O_j - \bar{O}]}{\sqrt{\sum_{j=1}^{n} [Y_j - \bar{Y}]^2 \sum_{j=1}^{n} [O_j - \bar{O}]^2}}
\]
where \(O\) is observed anomaly value, \(\bar{O}\) is mean of all the observed samples, \(Y\) is predicted anomaly value from models, \(\bar{Y}\) is mean of all the predicted samples, and \(n\) is number of samples.

Anomalies of the reproduced multi-year SSTs are obtained by subtracting their climatology. Anomalies are computed for each set of multi-year prediction with respect to an initial year (e.g. the lead year \(n = 1\), 2, ..., 10 corresponds to years 1960 + \(n\), 1965 + \(n\), ..., 2000 + \(n\), 2005 + \(n\) for the ten decades starting from 1960, 1965, ..., 2000, 2005). The simulated data is obtained from the core experimental set of GCMs (Taylor et al. 2012). Similarly, anomalies for the same decadal period are computed from the observational ERSST dataset. All the anomalies for a particular predicted year ‘\(n\)’ are compiled together to find the ACC values with a similar subset of observed anomalies. This is a standard approach recommended by the World Climate Research Programme (ICPO 2011) and is widely used in the assessment of decadal predictability (García-Serrano and Doblas-Reyes 2012, Kim et al. 2012).

2.4. Multivariate singular spectrum analysis and decadal predictions
MSSA was developed for extracting the dominant nonlinear modes of trend, periodicity and noise from a signal (Broomhead and King 1986, Golyandina et al. 2001). We implement MSSA to a multivariate dataset of interannual anomalies SSTs during March–May (MAM) and June–August (JJA) over 12 oceanic regions, with respect to their seasonal means. The analysis is applied to all the 12 time series together using a window length of 20 years. Similar to the decadal predictions made by the climate models, 10 year predictions of SSTs over the 12 oceanic regions are performed here from different initialized time series of 1901–1960, 1901–1965, ..., 1901–2005. MSSA extracts common signals from all the 12 SSTs together for a specific time length. The method further divides SST time series into three parts, viz., trend, harmonic oscillations and noise. The principle of separation is based on the fact that the eigenvectors represent the components graphically (Golyandina et al. 2001). We separate trends and oscillations from the noise in the time series for the decadal prediction of SSTs. The principles such as graphical representation, pairing and W-correlation matrix of eigenvectors are used to determine the extract nature of separated components. For detailed information about the extraction of signals, please see supplementary text S2. To further classify a particular eigenvector to trend, oscillations or noise, we apply fast Fourier transform (FFT) to understand the true nature of the component. From the above analysis, only trend and oscillatory reconstructed signals obtained for 12 SSTs, are summed together to perform the 10 year predictions.

For the decadal prediction of the newly reconstructed time series from MSSA (with trend and oscillatory signals only), we use linear recurrence relation (LRR) (Van Der Veen et al. 1993). The method tries to learn from the historical reconstructed trend and oscillatory modes of the time series. The predictive skills from the proposed model are further measured using ACC and compared with those of GCMs. We have provided more information regarding the LRR prediction technique in supplementary text S2.

We employ two statistical models for the decadal prediction of ISMR (as mentioned in the Introduction). In the first model, we employ SSA for the 10 year prediction of ISMR, which is purely data-driven. The model uses different initialized time series of ISMR, viz., 1901–1960, 1901–1965, ..., 1901–2005 to perform SSA predictions. The prediction skill of this model is further compared with the GCMs using ACC values for the predicted years ‘\(n\)’ with the observed data as the reference. Note that the ACC values will be determined using observed ISMR data. In the second model, ISMR is predicted from the reconstructed SSTs obtained from MSSA model. A regression model is trained using reconstructed SSTs over 12 oceanic regions for MAM and JJA SSTs as regressors and ISMR as predictand with different initialized time lengths (as in the earlier model). The 10 year predictions of SSTs obtained from the MSSA model are used for the decadal prediction of ISMR. We evaluate the performance of the proposed decadal prediction models with that of the GCMs.

3. Results
3.1. Nonlinear limits of predictability
We determine the nonlinear limits of predictability for 12 SST regions, which are associated with ISMR. The estimation of NLLEs considers observed monthly SST values during 1901–2010 over these oceanic regions. The results show that the highest attainable limits in the tropical region are \(\sim 11–12\) months and extratropical oceanic regions have \(\sim 4–5\) months (figure 1(b)). Niño4 and Niño3.4 regions show the maximum nonlinear limits of predictability among the SSTs. The results are found to be consistent with the nonlinear limits as obtained by Li and Ding.
(2013b). Barnston et al (2012) suggested that any good model could provide prediction skills up to 6–12 months for the most predictable component of SSTs, i.e. El Niño, which is consistent with our results regarding the limits of predictability. Considering all the regions together, the SST prediction system would forget the initial state of SSTs within the next 12 months (or 1 year) at any oceanic region.

### 3.2. Identifying trend and oscillation using MSSA

We perform decadal predictions of SSTs with different initialized time series using MSSA and compare their ACC skills with eight GCMs. We use interannual anomalies of SSTs during MAM and JJA, which are important for ISMR predictions. As we apply MSSA to the 12 SSTs for each season, the eigenvectors obtained from the analysis represent different components graphically. Supplementary figures S1 and S2 show the eigenvectors obtained from SSTs during MAM and JJA, among which the first set of vectors present the trends for both seasons (supplementary figures S1a and S2a). Other eigenvectors are representing other signals such as oscillations and noise. Here, we have used three methods to differentiate between oscillations and noise: (a) scatterplot between eigenvectors, (b) W-correlation matrix, and (c) FFT. The eigenvectors other than the trend components represent oscillations and noise. To differentiate between oscillations and noise, we use the scatterplot between two consecutive eigenvectors. A pair of eigenvectors representing oscillations will produce a polygon shape on a 2D scatter (supplementary figures S3(c) and (f), and supplementary figures S4(d) and (f)). In the method based on W-correlation matrix (supplementary figure S5), the principle involved is that well-separated components will have a small correlation. The badly separated components have higher correlation among them. The components with trend and periodicity cannot show higher correlation values. Moreover, the periodic signals are also phased at an angle of \( \pi/2 \). Hence the eigenvectors representing them would also have correlation values close to zero. Noisy components may show higher correlation values than the others. We further use FFT to confirm that the classifications of eigenvectors to trend, oscillations, and noise are correct. After performing FFT, the frequency-time graph shows a peak at time 0 for the reconstructed trends. The spectrum shows a single peak on the frequency-time graph for oscillatory signals. For noise, the FFT resembles a white spectrum with the existence of multiple peaks. We find that all three methods applied independently to the eigenvectors result in a similar classification. Detailed information regarding separation and categorization of these eigenvectors to trend, oscillations and noise is provided in supplementary text S2. After the identification and separation of all the different signal components, we combine trend and oscillatory components of SSTs for different initialized time series 1901–1960, 1965, 1970, ..., 2005 to generate a reconstructed time series. As for illustration, figure 2 shows the comparison between reconstructed SST time series (in blue) and observed SST time series (in red) over each oceanic region during MAM for the duration of 105 years (1901–2005). Trend and oscillations in the original time series at interannual, decadal and multidecadal time scales (clearly seen in extratropical NA, NP1, NP2 oceanic regions) are well captured in the reconstructed time series. Similarly, supplementary figure S6 shows a comparison of observed (in red) and reconstructed time series (in blue) of SSTs over 12 oceanic regions during JJA season, and the results are similar to those of MAM. The observed time series show larger amplitudes as they consist of noise along with trend and oscillations obtained from reconstruction using MSSA. The purpose of separating the trend and oscillations (periodicity) from the noisy signals is to perform the decadal predictions with the low frequency components, as they have the predictive potentials.

### 3.3. Decadal prediction skills of MSSA

After generating the reconstructed time series by combining the trend and periodic signals, we apply LRR to the newly reconstructed time series for the decadal prediction during MAM and JJA using MSSA. As for an example, for the decadal prediction with 1960 initial condition, we first generate the reconstructed time series for 1901–1960 by combining trend and oscillatory components and then apply LRR to this reconstructed time series for the prediction of the period 1961–1970. Now onwards, we will use the nomenclature MSSA to refer to this prediction model that considers MSSA followed by LRR. The multi-model ensemble (MME) is also produced by considering the mean of SST predictions from eight GCMs. ACC skill scores are obtained for the predicted SSTs from GCMs, MME and the MSSA model (figure 3), when compared to the observed SSTs. ACC values show that the predictive skill of MSSA are quite comparable to those of the GCMs. Interestingly, the SSTs predicted by both the MSSA and GCMs over AtlNiño and NA regions show the most promising decadal prediction skills among all the SST regions (figure 3). The higher skill of the models in predicting SSTs over these regions is well documented in literature (Smith et al 2007, Keenlyside et al 2008, Pohlmann et al 2009, van Oldenborgh et al 2012). The decadal prediction skills of MSSA over the BoB, Niño 4, NP1, SIO and WestIO regions are similar to those from GCMs and MME at different lead times. For the decadal prediction of 12 SSTs during JJA over the oceanic regions such as NA, Niño 3, Niño 4, NP1, NP2, SCS, SIO and WestIO, patterns of skill score of MSSA resemble well with those of the GCMs across all the lead times (supplementary figure S7). SSTs predicted over AtlNiño, BoB, NA and SIO regions from the MSSA model
and GCMs show positive ACC scores. Similarly, the ACC values from MSSA model over NA, SIO and WestIO show a positive skillset, which is in resemblance with those obtained from GCMs and MME (supplementary figure S7). For other regions, the ACC values obtained from both MSSA and the GCMs are around zero across different lead times. Similarity between the ACC skills as obtained from MSSA and GCMs (with MME) shows that the statistically retrieved trend and oscillations from SST time series are sufficient to produce decadal predictions with a skill similar to GCMs.

We investigate decadal prediction skills of models for the lead times of 1 year, 2–5 years and 6–9 years during MAM and JJA seasons (supplementary tables 1 and 2). The internal variability is the source of high skill at short lead times and the forced trend is responsible for skills at long lead times (Krueger and Von Storch 2011, Ho et al 2013). CanCM4 and BCC models largely show the ACC values decreasing with the increasing lead times for SST regions. However, MPI shows higher skills at 1 year and 6–9 years lead time than 2–5 years during MAM season (supplementary figure S8). The rest of the GCMs and MME do not have any specific characteristics to be pointed out. The skills for all the lead times obtained from MSSA predictions are comparable with those from GCMs (supplementary figures S8, S9). This shows the aggregated trend and oscillations in MSSA perform well at shorter and longer lead times.

3.4. Decadal prediction model for ISMR

Following the good performance of data-driven decadal prediction of SSTs using MSSA, we try the same procedure for ISMR. We collect the set of time series of 1901–1960, 65, 70,…, 2005 observed values of ISMR and apply SSA for 10 year prediction. This prediction model is purely data-driven, derived from the combined trend and oscillatory signals of observed ISMR. Figure 4 shows the skills score of predicted ISMR using the SSA method. We find that the ACC scores are comparable to the skills of eight GCMs and the MME. However, the skills are mostly negative or around zero for most of the lead years for all the models, SSA, GCMs and MME. We also implement a physics guided data-driven model, where we use the reconstructed SSTs obtained from the previous step as covariates for prediction of ISMR. The reconstructed

Figure 2. Observed timeseries (in red) and reconstructed time series using MSSA (in blue) for the SSTs in MAM during 1901–2005 over the oceanic regions: (a) AtlNiño, (b) BoB, (c) NA, (d) Niño3, (e) Niño4, (f) Niño3.4, (g) NP1, (h) NP2, (i) SCS, (j) SIO, (k) WestIO, (l) IOD.
SSTs from MSSA model over the 12 oceanic regions during MAM and JJA are used as predictors and the observed ISMR as predictand for the training of a linear regression model. The 10 year predictions of ISMR are made by applying the trained linear regression model to the predicted SST values of MAM and JJA from the MSSA model. As the selection of the covariates, SSTs over the 12 oceanic regions are based on the physical understanding of ISMR, the model can be termed as physics guided data-driven model. The skills of this decadal prediction algorithm for all the initialized sets of time series are further compared with those of eight GCMs and MME decadal outputs. We acknowledge that multicollinearity (high correlation) exists between the SST predictors. This results in an ill-formation of a regression model when the objective is to understand the dependent variable's sensitivity to an independent variable's perturbation. However, multicollinearity rarely affects the correctness of predicted values (Gujarati 2004). The ill-formation of the regression model may indeed result in overfitting. Figure 4 shows the comparison of decadal prediction skills of the physics guided data-driven model, purely data-driven model and eight GCMs with their MME in predicting ISMR. The physics-guided model shows higher skills with mostly positive correlation values indicating better performance than the GCMs (with their MME) and purely data-driven model.

4. Discussion

We have demonstrated that despite of the very high computational cost of GCMs, the decadal
performance skills of GCMs with their MME are generally poor with a very few exceptions. There are certain oceanic regions, viz., AtlNiño, BoB, NA, Niño 3, Niño 4, Niño 3.4, SIO and WestIO, over which the SSTs are predicted decadally by the GCMs with moderate skill, but no skills for many variables such as precipitation. This also questions the justification of spending huge computing time and associated energy for scientific activities related to decadal predictions without any fruitful results. With this background we have developed purely data-driven and physics guided data-driven models for the decadal prediction of one of the most complex processes, ISMR. We find that, with the requirement of negligible computing resources, the performance of the data-driven model is at par and the performances of physics guided data-driven model is better compared to those of GCMs up to the lead time of 6 years, after which the predictability of ISMR is lost.

To investigate the reason behind the improved performance by data-driven models, we revisit figure 2. The time series of the SSTs and the reconstructed plots show that the variability of SSTs over the AtlNiño, BoB, NA, SIO and WestIO regions are quite well explained by the trend and oscillations, with a minimum noise component. Hence, they are also well reproduced by the reconstructed time series by MSSA. These are the regions where the data-driven model shows high-performance skill for decadal predictions in terms of ACC score (figure 3). On the other hand, the time series of the SSTs over the regions Niño3, Niño 4, Niño3.4, IOD, SCS and NP1 have significant noise components, with random fluctuations (figure 2) and they are also not well predicted by MSSA (figure 3). It is also noteworthy that the skills of the majority of the GCMs follow similar patterns as of MSSA. This is expected because any model, irrespective of being driven by physics or data, will perform well for the variables that have more prominent signals of trend and oscillations and less noise.

The CMIP5 GCMs are observed to perform poorly for the decadal prediction of ISMR, despite the fact that the new generation models are able to reproduce the ENSO–ISMR coupling (Azad and Rajeevan 2016, Ramu et al 2018, Yun and Timmermann 2018, Pandey et al 2020). The CMIP5 models are in general good in terms of capturing the teleconnection of monsoon with ENSO and IOD (Roy and Kripalani 2019); but they fail to capture other important teleconnections such as with AtlNiño (Yadav et al 2018). Interestingly, the time series of SSTs over ENSO region and IOD are noisy (figure 2), and hence, they are not well predicted decadally by the GCMs. The signals from poorly predicted SSTs over the ENSO or IOD regions by CMIP5 GCMs, when transferred to ISMR through the atmosphere–ocean coupling, do not result in high decadal performance skill for ISMR. On the other hand, though models are good in predicting SSTs over the regions like AtlNiño, the signals are not getting mapped to the predicted ISMR as the performances of GCMs for such teleconnections are poor. In the MSSA model, the predictable signals from well predicted SSTs over certain oceanic regions (e.g. AtlNiño) are possibly transferred quite well to ISMR through the regression. Hence, the physics guided data-driven model outperforms the state of art GCMs up to the lead time of 6 years. Further to this, the physics guided data-driven model, for decadal predictions, considers the magnitude of a signal more than just a single initial state to initialize, unlike the GCMs, and tries to capture the historical patterns, signals and teleconnections from the long term datasets. Here, in the present study, we considered all the predictors together without dropping any correlated variables. The association or correlation between the predictors is not stationary and may change in an altered future climate. Hence, there is a possibility that dropping a predictor variable may result in a substantial loss of performance of the prediction (Braunisch et al 2013).

The present analysis is a proof of concept that the computationally inexpensive physics guided data-driven models have the potential to overcome many of the limitations of purely dynamic models. One of the limitations of the present model is that it does not explicitly consider the CO2 concentration, GHG emissions, anthropogenic aerosols, etc, though their impacts are probably implicitly being considered through the trend components of different variables. Extrapolation of trends is possibly acceptable when the prediction lead time is within 10 years; however, the long-term simulations need explicit consideration of different forcings possibly as covariates.

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

The work is financially supported by Department of Science and Technology Swarnajayanti Fellowship Scheme, through Project No. DST/SJF/E&ASA-01/2018-19; SB/SJF/2019-20/11, and Strategic Programs, Large Initiatives and Coordinated Action Enabler (SPLICE) and Climate Change Program through Project No. DST/CCP/GoE/140/2018. We acknowledge the World Climate Research Programme’s Working Group on Coupled Modelling, and thank the climate modelling groups (listed in Table 1 of this paper) for producing and making available their model output. The authors thank the ESGF for making the CMIP5 simulations freely available at https://esgf-node.llnl.gov/search/cmip5/. We acknowledge National Oceanic and Atmospheric Administration (NOAA) to freely
provide ERSST v3b data (www.ncdc.noaa.gov/data-access/marine-ocean-data/extended-reconstructed-sea-surface-temperature-erssst-v3b). We would also like to thank India Meteorological Department to provide high resolution rainfall data over India www.imdpune.gov.in/Clim_Pred_LRF_New/Grided_Data_Download.html (Accessed on 5 May 2021).

Contributions

S G conceived the idea and designed the problem. R S and S G developed the methodology. R S wrote the codes, performed the analysis. R S and S G analysed the results. R S and S G wrote the manuscript. R S prepared the figures.

Ethics declaration

The authors declare no competing interests.

ORCID iD

Subimal Ghosh https://orcid.org/0000-0002-5722-1440

References

Annamalai H, Hamilton K and Sperber K R 2007 The South Asian summer monsoon and its relationship with ENSO in the IPCC AR4 simulations J. Clim. 20 1071–92
Azad S and Rajeevan M 2016 Possible shift in the ENSO-Indian monsoon rainfall relationship under future global warming Sci. Rep. 6 1–6
Barnston A G, Tippett M K, L’Heureux M L, Li S and DeWitt D G 2012 Skill of real-time seasonal ENSO model predictions during 2002–11: is our capability increasing? Bull. Am. Meteorol. Soc. 93 631–51
Bollasina M A, Ming Y and Ramaswamy V 2011 Anthropogenic aerosols and the weakening of the South Asian summer monsoon Science 334 502–5
Branstator G, Teng H, Mehl G A, Kimoto M, Knight J R, Latif M and Rosati A 2012 Systematic estimates of initial-value decadal predictability for six AOGCMs J. Clim. 25 1827–46
Braunisch V, Coppejans G, Arlettaz R, Suchant R, Schmid H and Bollmann K 2013 Selecting from correlated climate codes, performed the analysis. R S and S G analysed the results. R S and S G wrote the manuscript. R S prepared the figures.

Ethics declaration

The authors declare no competing interests.

ORCID iD

Subimal Ghosh https://orcid.org/0000-0002-5722-1440

References

Annamalai H, Hamilton K and Sperber K R 2007 The South Asian summer monsoon and its relationship with ENSO in the IPCC AR4 simulations J. Clim. 20 1071–92
Azad S and Rajeevan M 2016 Possible shift in the ENSO-Indian monsoon rainfall relationship under future global warming Sci. Rep. 6 1–6
Barnston A G, Tippett M K, L’Heureux M L, Li S and DeWitt D G 2012 Skill of real-time seasonal ENSO model predictions during 2002–11: is our capability increasing? Bull. Am. Meteorol. Soc. 93 631–51
Bollasina M A, Ming Y and Ramaswamy V 2011 Anthropogenic aerosols and the weakening of the South Asian summer monsoon Science 334 502–5
Branstator G, Teng H, Mehl G A, Kimoto M, Knight J R, Latif M and Rosati A 2012 Systematic estimates of initial-value decadal predictability for six AOGCMs J. Clim. 25 1827–46
Braunisch V, Coppejans G, Arlettaz R, Suchant R, Schmid H and Bollmann K 2013 Selecting from correlated climate variables: a major source of uncertainty for predicting species distributions under climate change Ecography 36 971–83
Broomhead D S and King G P 1986 Extracting qualitative dynamics from experimental data Physica D 20 217–36
Buckley M W, DelSole T, Susan Lozier M and Li L 2019 Predictability of North Atlantic Sea surface temperature and upper-ocean heat content J. Clim. 32 3005–23
Chen B, Li J and Ding R 2006 Nonlinear local Lyapunov exponent and atmospheric predictability research Sci. China Ser. D: Earth Sci. 49 1111–20
Chikamoto Y et al 2013 An overview of decadal climate predictability in a multi-model ensemble by climate model MIROC Clim. Dyn. 40 1201–22
Collins M, Frame D, Sinha B and Wilson C 2002 How far ahead could we predict El Niño? Geophys. Res. Lett. 29 130–1–4
Constantinescu E M and Anitescu M 2013. Physics-based covariance models for Gaussian processes with multiple outputs Int. J. Uncertain. Quantification 3 47–71
Delsole T, Jia L and Tippett M K 2013 Decadal prediction of observed and simulated sea surface temperatures Geophys. Res. Lett. 40 2773–8
DelSole T and Tippett M K 2007 Predictability: recent insights from information theory Rev. Geophys. 45 RG4002
DelSole T and Tippett M K 2008 Predictable components and singular vectors J. Atmós. Sci. 65 1666–78
Devanand A, Roxy M K and Ghosh S 2018 Coupled land-atmosphere regional model reduces dry bias in Indian summer monsoon rainfall simulated by CSFv2 Geophys. Res. Lett. 45 2476–86
Ding R and Li J 2007 Nonlinear finite-time Lyapunov exponent and predictability Phys. Lett. A 364 396–400
Elhamod M, Bu J, Singh C, Redell M, Ghosh A, Podolskiy V, Lee W C and Karpatne A 2020 CoPhy-PGNN: learning physics-guided neural networks with competing loss functions for solving eigenvalue problems (arXiv-2007)
Enfield D B, Mantas-Nuñez A M and Trimble P J 2001 The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental U.S. Geophys. Res. Lett. 28 2077–80
Fan F, Mann M E, Lee S and Evans J L 2012 Future changes in the South Asian summer monsoon: an analysis of the CMIP3 multimodel projections J. Clim. 25 3909–28
Gadgil S 2003 The Indian monsoon and its variability Annu. Rev. Earth Planet. Sci. 31 429–67
Ganguly A R et al 2014 Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques Processes Geophys. 21 777–95
Garcia-Serrano J and Doblas-Reyes F J 2012 On the assessment of near-surface global temperature and North Atlantic multi-decadal variability in the ENSEMBLES decadal hindcasts Clim. Dyn. 39 2025–40
Goddard L, Hurrell J, Kirtman B, Murphy J, Stockdale M and Vera C 2012b Two Time scales for the price of one Bull. Am. Meteorol. Soc. 93 621–9
Golyandin N, Nekrutkin V and Zhigljavsky A 2001 Analysis of Time Series Structure: SSA and Related Techniques (Boca Raton, FL: CRC Press)
Guarneri D 2004 Basic Econometrics 4th edn (London: Macmillan)
Guo L, Turner A G and Highwood E J 2015 Impacts of 20th century aerosol emissions on the South Asian monsoon in the CMIP5 models Atmos. Chem. Phys. 15 6367–78
Hawkins E and Sutton R 2009 The potential to narrow uncertainty in regional climate predictions Bull. Am. Meteorol. Soc. 90 1095–107
Hidalgo H G and Alfaro E J 2015 Skill of CMIP5 climate models in reproducing 20th century basic climate features in central America Int. J. Climatol. 35 3397–421
Ho C K, Hawkins E, Shaffrey I and Underwood F M 2013 Statistical decadal predictions for sea surface temperatures: a benchmark for dynamical GCM predictions Clim. Dyn. 41 917–35
Huntingford C, Jeffers E S, Bonnall M B, Christensen H M, Lees T and Yang H 2019 Machine learning and artificial intelligence to aid climate change research and preparedness Environ. Res. Lett. 14 124007
ICPO 2011 Data and bias correction for decadal climate predictions International CLIVAR Project Office Publication Series 150 5 (available at: www.wcrp-climate.org/decadal/references/DCPP_Bias_Correction.pdf)
Jia X, Willard J, Karpatne A, Read J S, Zwart J A, Steinbach M and Kumar V 2020 Physics-guided machine learning for scientific discovery: an application in simulating lake ecosystems Proc. Nat. Acad. Sci. 117 13237–42
Karpatne A, Atluri G, Faghmous J H, Steinbach M, Banerjee A, Ganguly A, Shekhar S, Samatova N and Kumar V 2017a Theory-guided data science: a new paradigm for scientific discovery from data IEEE Trans. Knowl. Data Eng. 29 2318–31
Karpatne A, Watkins W, Read J and Kumar V 2017b Physics-guided neural networks (PGNN): an application in lake temperature modeling (arXiv:1710.11451)
Keenlyside N S and Ba J 2010 Prospects for decadal climate prediction. *Wiley Interdiscip. Rev. Clim. Change* 1 627–35
Keenlyside N S, Latif M, Jungclaus J, Kornbluh L and Roeckner E 2008 Advancing decadal-scale climate prediction in the North Atlantic sector *Nature* 453 84–8
Kim H-M, Webster J P and Carney J A 2012 Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts *Geophys. Res. Lett.* 39 L10703
Kim W M, Jhun J-G, Ha K-J and Kimoto M 2011 Decadal changes in climatological intraseasonal fluctuation of seasonal evolution of summer precipitation over the Korean Peninsula in the mid-1990s *Adv. Atmos. Sci.* 28 591–600
Knight J R, Folland C K and Scaife A A 2006 Climate impacts of the Atlantic multidecadal oscillation *Geophys. Res. Lett.* 33 L17706
Knutti R and Sedláček J 2013 Robustness and uncertainties in the new CMIP5 climate model projections *Nat. Clim. Change* 3 669–73
Kodra E, Bhatia U, Chatterjee S, Chen S and Ganguly A R 2020 Physics-guided probabilistic modeling of extreme precipitation under climate change *Sci. Rep.* 10 1–11
Kripalani R H, Oh J H, Kulkarni A, Sabade S S and Chaudhari H S 2007 South Asian summer monsoon precipitation variability: coupled climate model simulations and projections under IPCC A1A Theor. Appl. Climatol. 90 133–59
Krishnamurthy L and Krishnamurthy V 2014a Decadal scale oscillations and trend in the Indian monsoon rainfall *Clim. Dyn.* 43 319–31
Krishnamurthy L and Krishnamurthy V 2014b Influence of PDO on South Asian summer monsoon and monsoon–ENSO relation *Clim. Dyn.* 42 2397–410
Krishnamurthy L and Krishnamurthy V 2016 Teleconnections of Indian monsoon rainfall with AMO and Atlantic tripole *Clim. Dyn.* 46 2269–85
Krishnamurthy V and Shukla J 2008 Seasonal persistence and propagation of intraseasonal patterns over the Indian monsoon region. *Clim. Dyn.* 30 353–69
Krishnan R and Sugi M 2003 Pacific decadal oscillation and variability of the Indian summer monsoon rainfall *Clim. Dyn.* 21 233–42
Krueger O and Von Storch J S 2011 A simple empirical model for decadal climate prediction *J. Clim.* 24 1276–83
Kumar K K, Soman M K and Kumar K R 1995 Seasonal forecasting of Indian summer monsoon rainfall: a review *Weather* 50 449–67
Lau K M and Kim K M 2011 Comment on “Elevated heat pump” hypothesis for the aerosol–monsoon hydroclimate link: “Grounded” in observations? by S. Nigam and M. Bollasina *J. Geophys. Res.: Atmos.* 116 D17
Lean J L and Rind D H 2009b How will Earth’s surface temperature change in future decades? *Geophys. Res. Lett.* 36 L15708
Lee T C K, Zwiers F W, Zhang X and Tao M 2006b Evidence of decadal climate prediction skill resulting from changes in anthropogenic forcing *J. Clim.* 19 5305–18
Li J and Ding R 2013b Observational-spatial distribution of the predictability limit of monthly sea surface temperature in the global oceans *Int. J. Climatol.* 33 1936–47
Liermann F and Doblas-Reyes F J 2013b Decadal prediction of interannual tropical and North Pacific sea surface temperature *J. Geophys. Res.: Atmos.* 118 5913–22
Lorenz E N 1963b Deterministic nonperiodic flow *J. Atmos. Sci.* 20 130–41
Lorenz E N 1963 A study of the predictability of a 28-variable atmospheric model *Tellus* 17 321–33
Lu R, Dong B and Ding H 2006a Impact of the Atlantic multidecadal oscillation on the Asian summer monsoon *Geophys. Res. Lett.* 33 L24701
Malik A, Brönnimann S, Stickler A, Raible C C, Muthers S, Anet J, Rozanov E and Schmutz W 2017 Decadal to multi-decadal scale variability of Indian summer monsoon rainfall in the coupled ocean-atmosphere-chemistry climate model SOCOL-MPIOM *Clim. Dyn.* 49 3551–72
McPhaden M J, Zebiak S E and Glantz M H 2006 ENSO as an integrating concept in earth science *Science* 314 1740–5
Meehl G A et al. 2014 Decadal climate prediction: an update from the trenches *Bull. Am. Meteorol. Soc.* 95 243–67
Menon A, Levermann A, Schewe J, Lehmann J and Frieler K 2013 Consistent increase in Indian monsoon rainfall and its variability across CMIP-5 models *Earth Syst. Dyn.* 4 287–300
Mishra S K, Sahany S, Salunke P, Kang I S and Jain S 2018 Fidelity of CMIP5 multi-model mean in assessing Indian monsoon simulations *npj Clim. Atmos. Sci.* 1 1–8
Pai D S, Latha S, Rajeevan M, Sreejith O P, Sabthbai N S and Mukhopadhyay B 2014 Development of a new high spatial resolution (0.25° × 0.25°) Long period (1901–2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region MAUSAM 63 1–18
Pandey P, Dwivedi S, Goswami B N and Kucharski F 2020 A new perspective on ENSO-Indian summer monsoon rainfall relationship in a warming environment *Clim. Dyn.* 55 3307–26
Pawar S, Ahmed S E, San O, Rasheed A and Navon I M 2020 Long-term memory embedded nudging schemes for nonlinear data assimilation of geophysical *flows Phys. Fluids* 32 076006
Pohlmann H, Jungclaus J H, Köhl A, Stammer D and Marotzke J 2009 Initializing decadal climate predictions with the GECCO oceanic synthesis: effects on the North Atlantic *J. Clim.* 22 3926–38
Ramu D A, Chowdary J S, Ramakrishna S S V S and Kumar O S R U B 2018 Diversity in the representation of large-scale circulation associated with ENSO-Indian summer monsoon teleconnections in CMIP5 models *Theor. Appl. Climatol.* 132 645–78
Redolat D, Monjo R, Lopez-Bustins J A and Martín-Vide J 2019 Upper-level mediterranean oscillation index and seasonal variability of rainfall and temperature *Theor. Appl. Climatol.* 135 1059–77
Reichstein M, Camps-Alló G, Stevens B, Jung M, Denzler J, Carvalhais N and Prabhat 2019 Deep learning and process understanding for data-driven Earth system science *Nature* 566 195–204
Roxy M K, Ritika K, Terray P, Murtugudde R, Ashok K and Goswami B N 2015 Drying of Indian subcontinent by rapid Indian Ocean warming and a weakening land-sea thermal gradient *Nat. Commun.* 6 1–10
Roy I and Kripalani R H 2019 The role of natural factors (part 1): addressing on mechanism of different types of ENSO, related teleconnections and solar influence *Theor. Appl. Climatol.* 137 469–80
Salveari C T, Ramu Dandi A, Dhakate A, Salunke K, Mahapatra S and Rao S A 2013 Simulation of boreal summer intraseasonal oscillations in the latest CMIP5 coupled GCMs *J. Geophys. Res.: Atmos.* 118 4401–20
Saha A, Ghosh S, Sahana A S and Rao E P 2014 Failure of CMIP5 climate models in simulating post-1950 decreasing trend of Indian Ocean warming and a weakening land-sea thermal gradient *Nat. Commun.* 5 1–10
Sahastrabuddhe R, Ghosh S, Saha A and Murtugudde R 2019 A minimalistic seasonal prediction model for Indian monsoon based on spatial patterns of rainfall anomalies *Clim. Dyn.* 52 637–51
Sahastrabuddhe R, Ghosh S, Saha A and Murtugudde R 2019 A minimalistic seasonal prediction model for Indian monsoon based on spatial patterns of rainfall anomalies *Clim. Dyn.* 52 3663–81
Scher S 2018 Toward data-driven weather and climate forecasting: approximating a simple general circulation model with deep learning *Geophys. Res. Lett.* 45 12616–622
Smith D M, Casack S, Colman A W, Folland C K, Harris G R and Murphy J M 2007 Improved surface temperature prediction for the coming decade from a global climate model *Science* 317 796–9
Smith T M, Reynolds R W, Peterson T C and Lawrimore J 2008 Improvements to NOAA’s historical merged land–ocean surface temperature analysis (1880–2006) J. Clim. 21 2283–96

Solomon A et al 2011 Distinguishing the roles of natural and anthropogenically forced decadal climate variability and implications for prediction Bull. Am. Meteorol. Soc. 92 141–56

Sperber K R, Annamalai H, Kang I-S, Kitoh A, Moise A, Turner A, Wang B and Zhou T 2013 The Asian summer monsoon: an intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century Clim. Dyn. 41 2711–44

Suckling E B and Smith L A 2013 An evaluation of decadal probability forecasts from state-of-the-art climate models J. Clim. 26 9334–47

Sun C, Li J, Feng J and Xie F 2015 A decadal-scale teleconnection between the North Atlantic Oscillation and subtropical eastern Australian rainfall J. Clim. 28 1074–92

Sutton R T 2005 Ocean science: Atlantic Ocean forcing of North American and European summer climate Science 309 115–8

Taylor K E, Stouffer R J and Meehl G A 2012 An overview of CMIP5 and the experiment design Bull. Am. Meteorol. Soc. 93 485–98

Turner A G and Annamalai H 2012 Climate change and the South Asian summer monsoon Nat. Clim. Change 2 587–95

Turner A G and Slingo J M 2009 Uncertainties in future projections of extreme precipitation in the Indian monsoon region Atmos. Sci. Lett. 10 152–8

Van Der Veen A-J, Deprettere E F and Swindlehurst A L 1993 Subspace-based signal analysis using singular value decomposition Proc. IEEE 81 1277–308

van Oldenborgh G J, Doblas-Reyes F J, Wouters B and Hazeleger W 2012 Decadal prediction skill in a multi-model ensemble Clim. Dyn. 38 1263–80

Wang B, Xiang B, Li J, Webster P J, Rajeevan M N, Liu J and Hu K J 2015 Rethinking Indian monsoon rainfall prediction in the context of recent global warming Nat. Commun. 6 1–9

Webster P J, Magaña V O, Palmer T N, Shukla J, Tomas R A, Yanai M and Yasunari T 1998 Monsoons: processes, predictability, and the prospects for prediction J. Geophys. Res.: Oceans 103 14451–510

Wilcox I J, Highwood E J and Dunstone N J 2013 The influence of anthropogenic aerosol on multi-decadal variations of historical global climate Environ. Res. Lett. 8 024033

Willard J, Jia X, Xu S, Steinbach M and Kumar V 2020 Integrating physics-based modeling with machine learning: a survey (arXiv:2003.04919)

Yadav R K, Srinivas G and Chowdary J S 2018 Atlantic Niño modulation of the Indian summer monsoon through Asian jet npj Clim. Atmos. Sci. 1 1–11

Yun K-S and Timmermann A 2018 Decadal monsoon-ENSO relationships reexamined Geophys. Res. Lett. 45 2014–21

Zhang R and Delworth T L 2006 Impact of Atlantic multidecadal oscillations on India/Sahel rainfall and Atlantic hurricanes Geophys. Res. Lett. 33 L17712

Ziehmann C, Smith L A and Kurths J 2000 Localized Lyapunov exponents and the prediction of predictability Phys. Lett. A 271 237–51