Year-ahead predictability of South Asian Summer Monsoon precipitation

Environmental Research Letters

LETTER

Year-ahead predictability of South Asian Summer Monsoon precipitation

Nir Y Krakauer

Department of Civil Engineering and NOAA CREST, The City College of New York, New York, NY, 10031, United States of America

E-mail: mail@nirkrakauer.net

Keywords: seasonal forecasting, South Asian Summer Monsoon, sea-surface temperature, random forest, lasso

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
9 October 2018

REVISED
2 January 2019

ACCEPTED FOR PUBLICATION
21 January 2019

PUBLISHED
29 March 2019

Abstract

Since the South Asia Summer Monsoon is the main source of water for a densely cultivated and climate-sensitive region, its predictability has long been the target of research. This work estimates the predictability horizon of monsoon precipitation amount by systematically comparing statistical forecasts made using information from different lead times before the monsoon start. Linear and nonlinear prediction methods are considered that use the leading modes of the global sea surface temperature field to forecast monsoon-season (June–September) total precipitation on a 0.5° grid over South Asia, where each method is trained on data from 1901 to 1996 and evaluated on data from 1997 to 2017. Forecasts were found to outperform a climatology baseline up to at least 1 year ahead, with a nonlinear method (random forest) on average outperforming linear regression with group lasso, although with greater variability in skill across locations and years. Forecast performance measures (fractional reduction in root mean square error and information skill score) decreased with increasing lead time following exponential decay timescales of 5–12 months, depending on the performance measure and forecast method. Even at lead times of several years, there was some forecast skill compared to climatology, as a result of the impact of long-term climate change on monsoon precipitation. The results suggest that monsoon prediction is possible with longer lead times than generally attempted now.

1. Introduction

Densely populated South Asia, a leading agricultural area, derives most of its water supply from intense summer rainfall (figure 1). The rainy season is closely linked to the South Asia Summer Monsoon (SASM), in which moisture flows to the Indian subcontinent from the Arabian Sea and the Bay of Bengal, driven by the land-sea temperature contrast [1]. This is part of a summer monsoon zone that extends into East Asia and the northwest Tropical Pacific [2, 3].

Interannual variation in the amount and spatial distribution of summer precipitation in the SASM monsoon zone is considerable, with severe consequences for agricultural productivity and human well-being [4]. Prediction of monsoon precipitation has therefore been pursued for over a century, using predictors such as snow cover and air pressure patterns [5]. Over the 20th Century, the challenge of SASM precipitation prediction led to advances in statistical methods for multivariate time series analysis, and the discovery of large-scale interannual oscillation patterns including the El Niño Southern Oscillation (ENSO), North Atlantic Oscillation, North Pacific Oscillation, and Equatorial Indian Ocean Oscillation [6].

[7] provide a historical review of SASM seasonal prediction methods. Recent work on statistical approaches to SASM prediction includes an approach based on neural networks, with only monsoon precipitation from previous years as input [8]; seasonal prediction of regional precipitation based on sea surface temperature (SST) and air pressure tendencies [9]; one-month-ahead prediction based on ENSO and an Equatorial Indian Ocean oscillation index [10]; and prediction using springtime air pressure patterns over the Pacific [11]. Applications of atmosphere-ocean

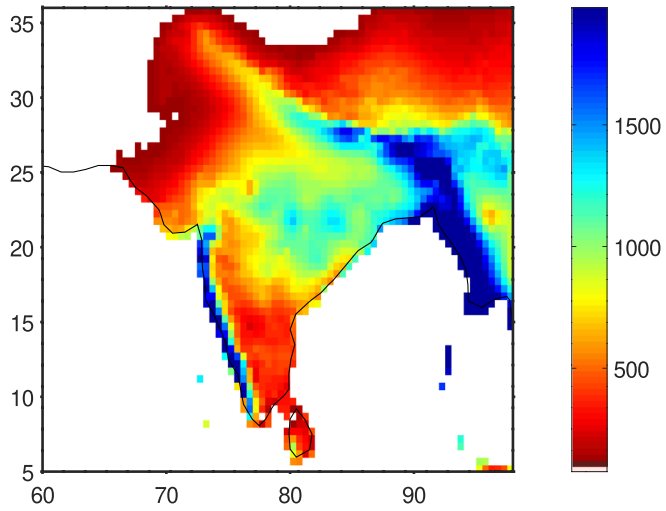

Figure 1. Mean monsoon-season (June–September) precipitation (mm) for South Asia, 1901–2016. Land grid cells where this was under 100 mm are left unshaded and were not included in the study domain.

dynamical model simulations to monsoon prediction also continue to be developed, with one study finding higher prediction skill of Indian summer monsoon rainfall at 3 month lead time despite poorer simulation of ENSO and Indian Ocean Dipole compared to shorter lead times [12], and another study showing a positive impact of higher model resolution on prediction skill for all-India summer monsoon rainfall [13]. However, despite enormous advances in dynamical simulation of climate in the past few decades, statistical methods remain competitive for forecasting at the seasonal timescale [14].

In recent years, improved statistical techniques have been developed and applied for predictive modeling where there are numerous possible predictors [14]. Such techniques may be linear or nonlinear in the predictor values. While many statistical models have been constructed to attempt to predict SASM precipitation, there are few systematic comparisons of predictive ability across forecast lead times. The main goal of the current study is to estimate the time horizon for predictability of the SASM precipitation field, using both linear and nonlinear state-of-the-art statistical methods. The basis for the predictions is taken to be the global SST field. SST reflects upper ocean heat content, which is recognized as the leading (though not only) contributor to weather and climate predictability on seasonal timescales [15]. Predictions are made for precipitation on a grid over South Asia, which may be more useful than predictions of single quantities such as regionwide precipitation indices or precipitation at single sites, as carried out by many of the previous studies.

2. Methods

2.1. Monsoon precipitation data

Spatially-resolved yearly accumulation of monsoon-season (June–September) precipitation over the South Asia region (5 to 36° N, 60 to 98° E; figure 1) was considered as the prediction target. Monthly precipitation gridded at 0.5° resolution for 1901–2017 was obtained from the publicly available University of East Anglia CRU-TS product (version 4.01), which interpolates observations from meteorological stations [16]. This product has been extensively intercompared with other products and observation datasets, with generally favorable results [17, 18]. Aggregated over India, the monsoon-season precipitation from this product compared well with the All-India Monsoon Rainfall Index, which is based on 36 representative stations [19], as obtained from the LDEO/IRI Climate Data Library [20]; the correlation coefficient between the two records over their period of overlap (1901–1998) was 0.933. There were 2215 grid cells with precipitation data over South Asia and an average of at least 100 mm per year in monsoon-season precipitation, covering 6.2 million square km (figure 1). Over most of this domain, the monsoon season accounted for more than half of annual precipitation, underscoring the importance of predicting it. The precipitation data were divided into training and test subsets: data from 1901 to 1996 were used for training monsoon prediction models, while 1997–2017 data were used for evaluating model performance.

2.2. Predictor data

Monthly global SST fields were obtained from the Hadley Centre's HadISST product (version 1.1), which provides 1° spatial resolution since 1870 [21]. This product has been used extensively for purposes such as assessing teleconnections with precipitation and temperature extremes [22] and evaluating predictions of SST [23]. Grid cells with sea ice in this product were filled in with an SST of −1.8° C. Then, singular value decomposition (SVD) [24, 25] was performed on SST

from the training period, with grid cells weighted by their surface areas, to find dominant modes of interannual SST variability. Approximately 40 of the leading SVD singular vectors (with the exact number depending on the lead time selected), representing at least 90% of the interannual SST variance, were retained as the potential predictors of monsoon precipitation. Lead times from 0 to 48 months were considered for forecasting. A lead time of 0 months, for example, means that the June–September precipitation was forecast using June SST; for a lead time of 6 months, the previous December's SST was used for prediction.

2.3. Prediction models

2.3.1. Linear model: least absolute shrinkage and selection operator (lasso)

Given training data on monsoon precipitation anomalies as an m by n matrix $\mathbf{Y}$, where $m = 96$ is the number of years of training data and $n = 2215$ the number of locations, and predictor values for each year as an m by k matrix $\mathbf{X}$, where $k \approx 40$ is the number of predictors (SST modes), a linear model would take the form $\mathbf{Y} \sim \mathbf{XB}$ (assuming that the columns of $\mathbf{X}$ and $\mathbf{Y}$ have had their means over the training period subtracted so that no explicit intercept term is needed).

In this problem, the dimension k of the predictor space is relatively large compared to the number of training instances m. With no constraints on the k by n coefficient matrix $\mathbf{B}$, its estimation, for example by least squares, is unduly sensitive to noise in the data and does not enable accurate prediction of precipitation.

The lasso tends to be effective for prediction in a context where, out of the many possible predictors, a small to moderate number are moderately associated with the predicand [26]. Specifically, the group-lasso variant [27, 28] is adopted, which is suitable for multivariate regression. In this method, $\mathbf{B}$ is chosen to minimize the cost function

$$\frac{1}{2}||\mathbf{Y} - \mathbf{XB}||_{\mathrm{F}}^2 + \lambda \sum_{i=1}^{k} ||B_i||_2, \qquad (1)$$

where the first term is proportional to the residual sum of squares and the second, penalty, term is proportional to the sum of the vector two-norms of the rows of $\mathbf{B}$. As the non-negative regularization parameter λ grows, the optimal $\mathbf{B}$ has fewer rows with non-zero elements, corresponding to fewer predictors included in the regression model. For large enough λ, no predictors are kept ($\mathbf{B}$ has only zeros) and the prediction of the regression model is therefore equal to the mean of the training-period predicand values.

The value of λ was chosen by cross validation [29, 30] to minimize the prediction mean square error. In the cross validation, the training dataset was divided into six equal-length segments, and each sixth was successively predicted using the remaining segments under different values of λ. Given λ, $\mathbf{B}$ was found by minimizing the cost function (1) using the coordinate descent method, as implemented in the R `glmnet` package [31–34].

Initial experimentation compared the lasso approach to several others: reduced-rank regression, where the matrix $\mathbf{B}$ is constrained to have some low rank r [35]; ridge regression [36], where the penalty term in the cost function is proportional to $||\mathbf{B}||_{\mathrm{F}}$ and the coefficients of $\mathbf{B}$ are therefore shrunk toward zero but not specifically by row [37, 38]; or a combination of reduced rank with the ridge regression penalty [39]. However, for these approaches, cross-validation usually selected regularization parameters r, λ that result in $\mathbf{B}$ being all zeros and therefore not yielding a useful prediction. The lasso penalty was therefore chosen as apparently the most effective in this context. The lasso method has previously been applied, for example, to predict land climate using ocean climate quantities [40], study the influence of weather on fruit yields [41], and reconstruct temperature fields from proxy data [42].

2.3.2. Nonlinear model: random forest (RF)

The nonlinear model considered was regression RF. This generates an ensemble (forest) of regression trees to predict the training data points using the SST modes, latitude, longitude, year, and mean precipitation for the location as predictors. The predicted value is then the mean across the regression tree ensemble [43]. RF is robust to the presence of correlated or unhelpful predictors [44] and has been successfully used for many earth science applications, generally comparing favorably with linear and other nonlinear methods [45–52]. The implementation in the R `randomForest` package [53] was used. All the parameter values were kept at the package default except for the number of points randomly sampled and used to fit each tree, which was reduced to 20% of the total number mn, or about 40 000, to limit computation time.

2.3.3. Baseline model: climatology

Forecast method skill—here, the skill of the linear and nonlinear models, which both use the same SST modes as predictors—is measured relative to some 'no-skill' baseline forecast [54]. The baseline model here was one where the forecast for each grid point and year in the test period is simply the average precipitation for that point over the training period (climatology).

2.4. Prediction skill measures and visualization of forecast performance

2.4.1. Reduction in root mean square error (RMSE)

Each of the models considered was used to generate predictions for the test monsoon-season precipitation data (precipitation at 2215 grid points for each of

21 years). The deterministic skill measure adopted was based on the RMSE between the predicted and actual values:

$$\mathrm{RMSE} = \sqrt{\frac{1}{m'}\sum_{i=1}^{m'}\frac{\sum_{j=1}^{n} a_j (F_{i,j} - P_{i,j})^2}{\sum_{j=1}^{n} a_j}}, \qquad (2)$$

where $F_{i,j}$ is the predicted precipitation for each test year and grid cell, $P_{i,j}$ is the actual precipitation from CRU-TS, a_j is the grid-cell area, and $m' = 21$ is the number of years in the test period. This was compared to the RMSE of the baseline climatology forecast. The fractional reduction in RMSE (rRMSE) for the linear or nonlinear model compared to climatology was computed as

$$\mathrm{rRMSE} = 1 - \frac{\mathrm{RMSE}_{\mathrm{for}}}{\mathrm{RMSE}_{\mathrm{clim}}}, \qquad (3)$$

where $\mathrm{RMSE}_{\mathrm{for}}$ is the RMSE calculated for the linear or nonlinear model and $\mathrm{RMSE}_{\mathrm{clim}}$ that for the climatology.

Confidence intervals for rRMSE were computed using the *t* test for the time series of forecast-climatology differences in mean square error over each of the 21 test years. The standard assumptions for the *t* test are that the time series tested is homoskedastic and has no autocorrelation. To assess the impacts of autocorrelation in the time series on the confidence intervals, confidence intervals were also calculated using the Newey–West method, which adjusts the *t* test standard error based on estimators for the time series heteroskedasticity and autocorrelation [55–57].

2.4.2. Probabilistic prediction and information skill score (ISS)

The point predictions generated from linear or nonlinear regression models were also converted to probabilistic predictions that are often of greater usefulness in decisionmaking. Similar to current operational forecasts such as those from the South Asian Climate Outlook Forum [58], the probabilities generated were those for placing in each tercile of precipitation from the training period. The point predictions were converted to per-tercile probabilities that minimized the Kullback–Leibler divergence from an equal-chances climatology probability distribution while having the mean indicated by the point prediction (a 'maximum entropy' approach [59, 60]), but with the tercile probabilities all constrained to the range $25\% \leqslant p \leqslant 50\%$.

The probabilistic skill measure adopted was mean ISS relative to an equal-chances climatology probabilistic forecast (of 1/3 per tercile). The ISS is based on the negative log likelihood of the actual outcome tercile under the forecast, normalized so that the climatology forecast has ISS of 0 and a perfect forecast that always predicts the observed tercile would have ISS of 1 [54]. The *t* test was used to compute confidence intervals for ISS analogous to those for rRMSE.

2.4.3. Skill as a function of lead time

The skill scores rRMSE and ISS were thus found for each prediction method (linear and nonlinear) and lead time (0–48 months). An exponential decay function was fitted to the rRMSE or ISS versus lead time relationship using least squares to generate a smoothed representation of the expected relationship of lead time to prediction skill and help estimate the predictability horizon for the monsoon precipitation. For each prediction method, the form of the function was

$$S(t) = a + be^{-t/\tau}, \qquad (4)$$

where $S(t)$ is the smoothed prediction skill (rRMSE or ISS), a is the estimated skill at long lead times ($t \gg \tau$), $a + b$ is the estimated skill at zero lead time, and τ is a decay timescale for the prediction skill.

2.4.4. Predictive SST modes and forecast skill mapping

For linear regression, correlating specific SST patterns with precipitation patterns is straightforward. The first singular vectors from singular-value decomposition of the coefficient matrix **B** from the lasso regression at 0 and 12 month lead times were mapped to show the SST mode associated with the most monsoon precipitation variability at those lead times and the corresponding precipitation response spatial pattern. For nonlinear regression methods such as RF, depicting the specific SST configurations correlated with precipitation responses is more difficult, but the rRMSE skill score for each grid point was mapped for the nonlinear regression as well as for the linear regression to show where each method has predictive skill. The skill scores were based on all lead times from 0 to 12 months, to reduce fluctuations due to small sample size, and pointwise confidence intervals for them were calculated using the *t* test for the difference of mean square error from that of a climatology forecast across years and lead times.

2.5. Sensitivity analyses

In order to better understand the behavior of the prediction methods, a number of variants were tested for which prediction models were fitted and skill scores computed. Details and results from these analyses are given in supplementary material.

3. Results

3.1. Skill scores by lead time

The nonlinear SST-based forecast was able to reduce RMSE relative to climatology at all lead times from 0 to 17 months, by an average of 1.9% (figure 2). The fitted exponential decay curve, which smooths out

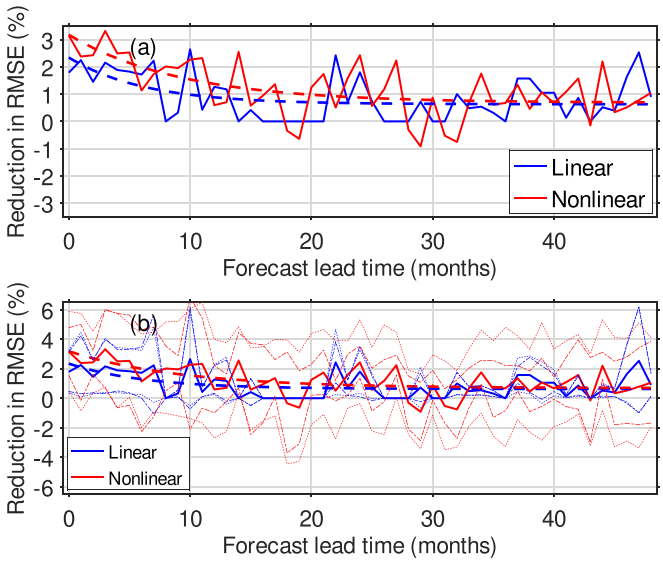

Figure 2. (a) Prediction skill for South Asia monsoon-season precipitation, quantified as reduction in root mean square error (rRMSE) relative to a forecast based on climatology, for two methods using the sea surface temperature (SST) field at different lead times as the predictor: linear regression with lasso, and nonlinear random forest regression. The thick dashed lines are exponential fits to the RMSE reduction as a function of lead time. (b) Same as (a), but with dotted lines showing 95% confidence intervals of rRMSE based on between-year variability of the forecast methods' mean square error, and with the vertical axis stretched to show the confidence ranges. The dashed–dotted lines show 95% confidence intervals when the Newey–West method is used to account for autocorrelation in the series of yearly forecast-skill differences.

fluctuations in the rRMSE between adjacent lead times due to sampling variability, gave an estimate of the expected rRMSE at zero lead as 3.2%, declining with an *e*-folding timescale of 9 months to 1.5% at 12 month lead and 0.7% at 48 month lead (figure 2). The linear SST-based forecast outperformed climatology by an average of 1.2% over 0 to 17 month lead times, declining with an *e*-folding timescale of 6 months from 2.3% at zero lag to 0.9% at 12 month lead and 0.6% at 48 months.

Although the linear method produced forecasts with on average somewhat less rRMSE than the nonlinear method, its results fluctuated less than the nonlinear one between adjacent leads, and showed narrower confidence intervals based on between-year variability in performance (figure 2(b)). At some lead times, particularly the longer ones, rRMSE for the linear method was exactly zero because cross-validation produced a large enough λ (equation (1)) that the coefficient matrix **B** had all zeros and the regression estimate was therefore simply the training-period mean precipitation, which was also the climatology forecast. The Newey–West confidence intervals for the linear regression rRMSE were almost the same as with the standard *t* test, suggesting little year-to-year autocorrelation of the skill. For the nonlinear regression, the Newey–West confidence intervals were often narrower than the standard ones, implying some negative year-to-year autocorrelation and resulting in more lead times for which the rRMSE 95% confidence interval was all above zero and the skill score therefore significantly greater than zero (at the $p = 0.025$ level, using a one-tailed *t* test) (figure 2(b)).

The probabilistic tercile forecasts for grid-scale monsoon-season precipitation also showed skill, as measured by ISS relative to climatology. Forecasts derived from the nonlinear method showed positive ISS for all leads up to 11 months (figure 3). Mean ISS over those leads was smaller for the linear method (0.7%) than for the nonlinear one (1.4%). The exponential fits to ISS as a function of lag were, for the nonlinear method, 2.1% at zero lead, falling with a 12 month decay timescale to 0.6% at 12 month lead and −0.3% at 48 months, and, for the linear method, 1.3% at zero lead, falling with a 5 month timescale to 0.4% at 12 month lead and 0.3% at 48 months. Although the nonlinear method showed higher ISS than the linear one for the shorter leads of up to a year, the linear method maintained a more consistently positive ISS at longer lead times, whereas the nonlinear method average ISS was negative at those leads (figure 3). This is consistent with ISS being more sensitive than deterministic skill scores like rRMSE to variability in predictive skill [54], which was more pronounced for the nonlinear than for the linear method.

3.2. Predictability patterns and skill score maps

At zero lead, the leading SST mode correlated with SASM precipitation included warm conditions in the eastern Tropical Pacific along with cool conditions in the northwest and southwest Pacific (figure 4(a)). The

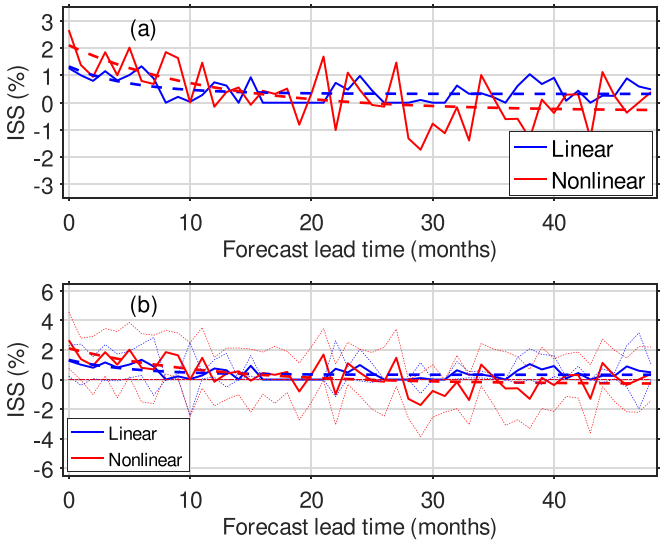

Figure 3. (a) Probabilistic prediction skill for South Asia monsoon-season precipitation, quantified as information skill score (ISS) relative to a forecast based on climatology, for two methods using the sea surface temperature (SST) field at different lead times as the predictor: linear regression with lasso, and nonlinear random forest regression. The thick dashed lines are exponential fits to the RMSE reduction as a function of lead time. (b) Same as (a), but with dotted lines showing 95% confidence intervals of rRMSE based on between-year variability of the forecast methods' mean square error, and with the vertical axis stretched to show the confidence ranges.

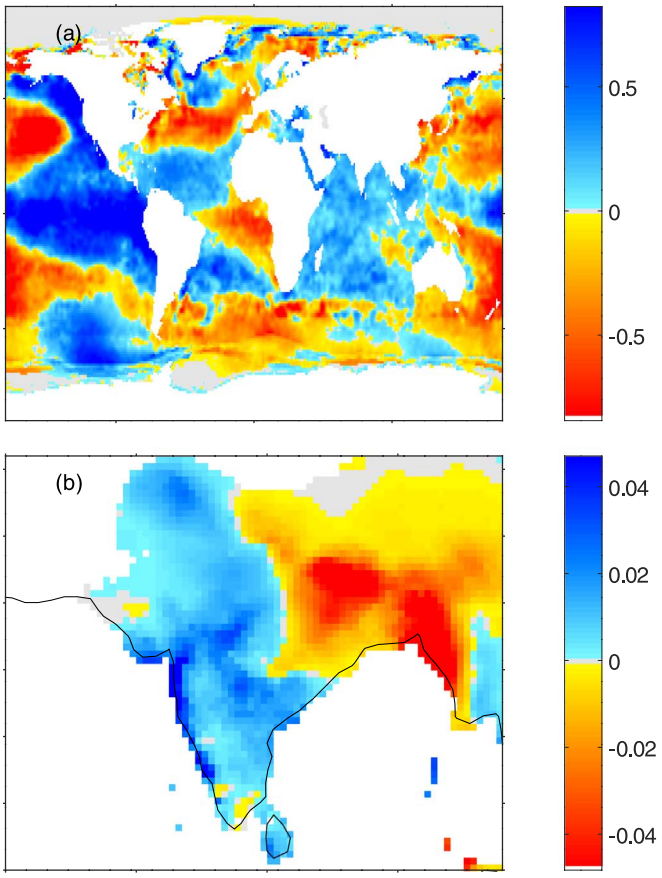

Figure 4. (a) June sea surface temperature most closely associated with South Asia monsoon-season (June–September) precipitation, as found using singular value decomposition from the lasso regression. (b) The associated South Asia monsoon-season precipitation response. The numerical scalings are arbitrary.

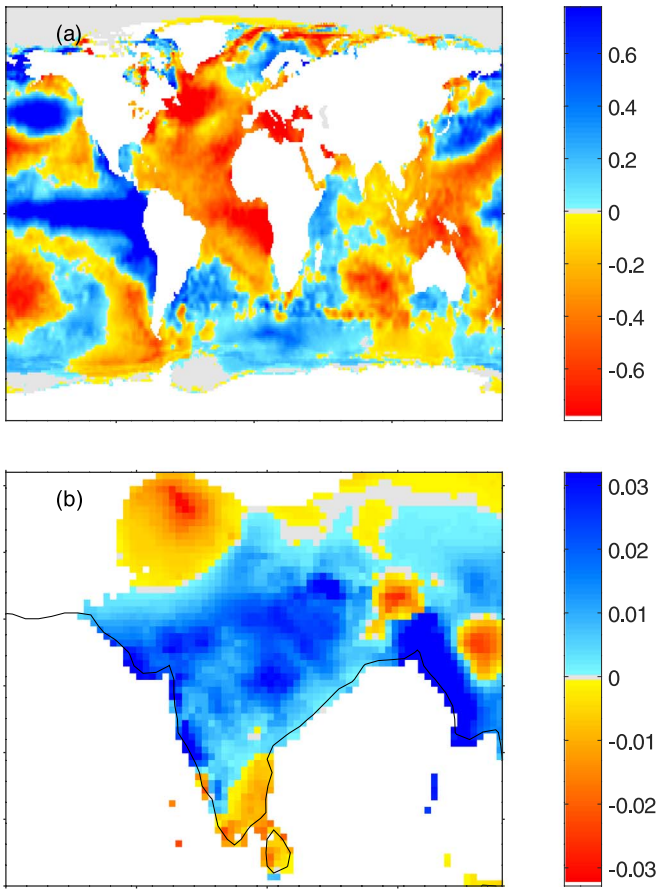

Figure 5. (a) Previous-year June sea surface temperature most closely associated with South Asia monsoon-season (June–September) precipitation, as found using singular value decomposition from the lasso regression. (b) The associated South Asia monsoon-season precipitation response. The numerical scalings are arbitrary.

corresponding precipitation response had an east–west contrast over the SASM region, with wet conditions in eastern and southern India along with dry conditions in the Ganges basin (figure 4(b)). At 12 month lead, the leading SST mode showed warm conditions along a narrower region of the Equatorial Pacific, along with cool conditions in most of the Atlantic (figure 5(a)). The precipitation response showed wet conditions over most of the SASM region, and dry ones on its southern and northwest margins (figure 5(b)). These SST patterns are weakly correlated with those associated with ENSO, with correlations of around −0.2 between their time series at both lead times and that of SOI.

To visualize what parts of South Asia show predictability for monsoon-season precipitation, the mean rRMSE, averaged over the 0–12 month lags to reduce fluctuations, was mapped for the linear and nonlinear forecasts. The two methods show similar though not identical regions with significant positive skill, particularly over the Ganges basin and Deccan Plateau, while forecasts had negative rRMSE (i.e. were less effective than the climatology forecast) at some points near the region's eastern and southern edges (e.g. lower Indus basin and Sri Lanka, respectively) (figure 6). Compared with the linear forecast (figure 6(a)), the nonlinear one (figure 6(b)) shows more positive rRMSE on average and in the areas where the linear forecast has skill, but also more negative rRMSE where the linear forecast has near-zero or negative rRMSE, indicating a less consistent forecast.

4. Discussion

Skill in prediction of monsoon-season precipitation over South Asia from the global SST field tended to decrease with increasing lead time, as expected, but remained consistently positive even over a year in advance, as measured by rRMSE relative to climatology (figure 2). This long-term predictability of precipitation could be valuable, for example, for water resource planners. Prediction skill relative to a forecast based on the 1901–1996 climatology was positive on average even at the longest lead times of 4 years, due to change in expected monsoon patterns associated with global warming. This phenomenon of climate-change-related seasonal to interannual prediction skill being comparable to that achievable by considering transient predictors is also found in seasonal forecasting studies from other regions [61–64], and means that

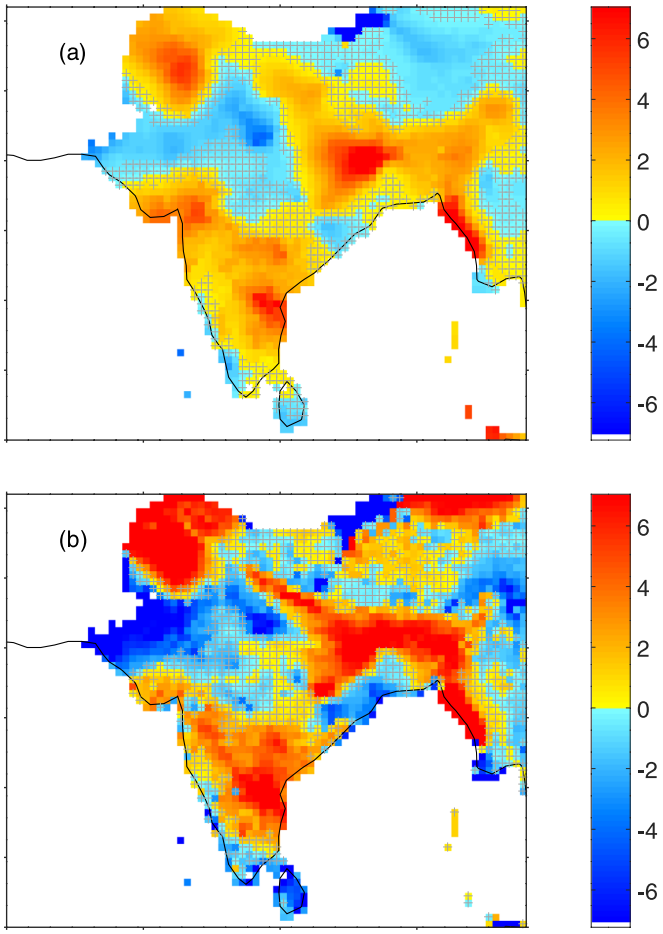

Figure 6. Percentage reduction in root mean square error for forecasts of monsoon-season precipitation based on sea surface temperature 0–12 months before, compared to a forecast based on climatology. (a) Linear-regression forecast model. (b) Nonlinear-regression forecast model. Cross-hatching shows areas where this reduction is not significantly different from zero at the 5% confidence level.

the baseline climatology period needs to be carefully defined in order to properly compare the skillfulness of different forecast methods.

Both a linear forecast method (lasso) and a nonlinear method (RF) showed skill in predicting monsoon-season precipitation. The nonlinear method showed higher skill on average but greater variability in skill across years and locations. Configurations of these methods and related ones were not evaluated exhaustively, so it cannot be concluded that these are the best possible methods or that nonlinear methods will in general outperform linear ones. The similarity in geographic distribution between the skill score spatial patterns (figure 6) suggests that both methods are exploiting similar associations of precipitation with the SST field. The sensitivity tests presented in the supplemental material suggest that each method may have its own strengths—for example, the nonlinear method appears more robust to the removal of the global warming component of SST evolution (figure S3 available online at stacks.iop.org/ERL/14/044006/mmedia), while the linear method does better when fewer SST modes are provided as predictors (figure S2). That detrending SSTs by removing the component linearly correlated with SOI reduced but did not eliminate skill (figure S3) is consistent with studies that have found that SST and pressure patterns outside the main ENSO region of the eastern Equatorial Pacific have become increasingly important predictors for Indian summer monsoon rainfall in recent decades [11, 12].

There are a number of approaches that could be evaluated for further improving monsoon precipitation forecast skill. Parameters in lasso and RF, here generally left at default values, could be systematically calibrated, as could other factors such as the number of SST modes used as predictors. Additional possible predictors, such as snow cover, soil moisture patterns, and atmospheric pressure patterns [65–70], could be added to see if their inclusion yields improvements over only using SSTs, particularly for shorter lead times of under a year. Prediction over the South Asia domain could be compared to predicting over smaller or larger (e.g. continental or

global) domains. Prediction could be based on the evolution of the SST field across months, rather than only on a single month's field as done here. For example, [71] found that the difference between spring and winter values of ENSO indices was a better predictor of monsoon-season irrigation requirement for a district in India than the winter or spring index values themselves, and [72, 73] found that monsoon precipitation in Nepal shows interannual autocorrelation and correlates with the Pacific Quasi-Decadal Oscillation lagged 2 years. Numerical weather prediction model outputs, which are known to have skill in monsoon prediction at least over lags of up to a few months [74], could be added as predictors in the forecast model.

It may be possible to improve the performance of these probabilistic tercile forecasts by refining the method used to derive them from the predictions of the linear or nonlinear SST-based forecast models. The tercile probabilities could also be predicted directly rather than estimated indirectly from the deterministic forecast, through for example quantile linear regression [75] and quantile RF [45, 76–78].

5. Conclusion

By considering linear and nonlinear forecast methods using SST modes as predictors, prediction skill for spatially distributed monsoon-period precipitation over South Asia was found to decay with a timescale of 5–12 months, but with residual skill at several-year lead times due to long-term climate trends. While these methods are not definitive and could likely be further improved, the present findings suggest that South Asia monsoon-period precipitation can be predicted with longer lead time than the subseasonal to seasonal leads usually attempted now.

Acknowledgments

The author gratefully acknowledges support from NOAA under grants NA16SEC4810008 and NA15OAR4310080 and by the United States Agency for International Development (USAID) under the US-Pakistan Centers for Advanced Studies in Water. All statements made are the views of the author and not the opinions of the funding agency or the US government.

ORCID iDs

Nir Y Krakauer https://orcid.org/0000-0002-4926-5427

References

[1] Krishnamurti T N 1985 Summer monsoon experiment—a review *Mon. Weather Rev.* **113** 1590–626
[2] Wang B and Ho L 2002 Rainy season of the Asian-Pacific summer monsoon *J. Clim.* **15** 386–98
[3] Day J A, Fung I and Risi C 2015 Coupling of South and East Asian monsoon precipitation in July–August *J. Clim.* **28** 4330–56
[4] Cook E R, Anchukaitis K J, Buckley B M, D'Arrigo R D, Jacoby G C and Wright W E 2010 Asian monsoon failure and megadrought during the last millennium *Science* **328** 486–9
[5] Normand C 1953 Monsoon seasonal forecasting *Q. J. R. Meteorol. Soc.* **79** 463–73
[6] Katz R W 2002 Sir Gilbert Walker and a connection between El Niño and statistics *Stat. Sci.* **17** 97–112
[7] Kumar K K, Soman M K and Kumar R K 1995 Seasonal forecasting of Indian summer monsoon rainfall: a review *Weather* **50** 449–67
[8] Dash Y, Mishra S K, Sahany S and Panigrahi B K 2018 Indian summer monsoon rainfall prediction: a comparison of iterative and non-iterative approaches *Appl. Soft Comput.* **70** 1122–34
[9] Li J, Wang B and Yang Y-M 2017 Retrospective seasonal prediction of summer monsoon rainfall over West Central and Peninsular India in the past 142 years *Clim. Dyn.* **48** 2581–96
[10] Surendran S, Gadgil S, Francis P A and Rajeevan M 2015 Prediction of Indian rainfall during the summer monsoon season on the basis of links with equatorial Pacific and Indian Ocean climate indices *Environ. Res. Lett.* **10** 094004
[11] Wang B, Xiang B, Li J, Webster P J, Rajeevan M N, Liu J and Ha K-J 2015 Rethinking Indian monsoon rainfall prediction in the context of recent global warming *Nat. Commun.* **6** 7154
[12] Pokhrel S, Saha S K, Dhakate A, Rahman H, Chaudhari H S, Salunke K, Hazra A, Sujith K and Sikka D R 2016 Seasonal prediction of Indian summer monsoon rainfall in NCEP CFSv2: forecast and predictability error *Clim. Dyn.* **46** 2305–26
[13] Ramu D A, Sabeerali C T, Chattopadhyay R, Rao D N, George G, Dhakate A R, Salunke K, Srivastava A and Rao S A 2016 Indian summer monsoon rainfall simulation and prediction skill in the CFSv2 coupled model: impact of atmospheric horizontal resolution *J. Geophys. Res.: Atmos.* **121** 2205–21
[14] Cohen J, Coumou D, Hwang J, Mackey L, Orenstein P, Totz S and Tziperman E 2018 S2S reboot: an argument for greater inclusion of machine learning in subseasonal to seasonal forecasts *Wiley Interdiscip. Rev.: Clim. Change* **0** e00567
[15] Committee on Developing a U.S. Research Agenda to Advance Subseasonal to Seasonal Forecasting, Board on Atmospheric Sciences and Climate, Ocean Studies Board, Division on Earth and Life Studies, and National Academies of Sciences, Engineering, and Medicine 2016 *Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts* (Washington, DC: National Academies Press)
[16] Harris I, Jones P D, Osborn T J and Lister D H 2014 Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 Dataset *Int. J. Climatol.* **34** 623–42
[17] Los S O 2015 Testing gridded land precipitation data and precipitation and runoff reanalyses (1982–2010) between 45° S and 45° N with normalised difference vegetation index data *Hydrol. Earth Syst. Sci.* **19** 1713–25
[18] Beck H E, Vergopolan N, Pan M, Levizzani V, van Dijk A I J M, Weedon G P, Brocca L, Pappenberger F, Huffman G J and Wood E F 2017 Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling *Hydrol. Earth Syst. Sci.* **21** 6201–17
[19] Sontakke N A, Pant G B and Singh N 1993 Construction of all-India summer monsoon rainfall series for the period 1844–1991 *J. Clim.* **6** 1807–11
[20] Blumenthal M B, Bell M, del Corral J, Cousin R and Khomyakov I 2014 IRI Data Library: enhancing accessibility of climate knowledge *Earth Perspect.* **1** 1–12
[21] Rayner N A, Parker D E, Horton E B, Folland C K, Alexander L V, Rowell D P, Kent E C and Kaplan A 2003 Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century *J. Geophys. Res.* **108** 4407

[22] Alexander L V, Uotila P and Nicholls N 2009 Influence of sea surface temperature variability on global temperature and precipitation extremes *J. Geophys. Res.* **114** D18116

[23] Ho C K, Hawkins E, Shaffrey L and Underwood F M 2012 Statistical decadal predictions for sea surface temperatures: a benchmark for dynamical GCM predictions *Clim. Dyn.* **41** 917–35

[24] Klema V and Laub A 1980 The singular value decomposition: its computation and some applications *IEEE Trans. Autom. Control* **25** 164–76

[25] Bretherton C S, Smith C and Wallace J M 1992 An intercomparison of methods for finding coupled patterns in climate data *J. Clim.* **5** 541–60

[26] Tibshirani R 1996 Regression shrinkage and selection via the lasso *J. R. Stat. Soc.* B **58** 267–88

[27] Yuan M and Lin Y 2006 Model selection and estimation in regression with grouped variables *J. R. Stat. Soc.* B **68** 49–67

[28] Breheny P and Huang J 2015 Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors *Stat. Comput.* **25** 173–87

[29] Efron B and Gong G 1983 A leisurely look at the bootstrap, the jackknife, and cross-validation *Am. Stat.* **37** 36–48

[30] Shao J 1993 Linear model selection by cross-validation *J. Am. Stat. Assoc.* **88** 486–94

[31] Friedman J, Hastie T and Tibshirani R 2010 Regularization paths for generalized linear models via coordinate descent *J. Stat. Softw.* **33** 1–22

[32] Breheny P and Huang J 2011 Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection *Ann. Appl. Stat.* **5** 232–53

[33] Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J and Tibshirani R J 2012 Strong rules for discarding predictors in lasso-type problems *J. R. Stat. Soc.* B **74** 245–66

[34] Simon N, Friedman J and Hastie T 2013 A blockwise descent algorithm for group-penalized multiresponse and multinomial regression arXiv:1311.6529

[35] Reinsel G C and Velu R P 1998 *Multivariate Reduced-Rank Regression : Theory and Applications* (*Lecture Notes in Statistics*) vol 136 (Berlin: Springer)

[36] Hoerl A E and Kennard R W 1970 Ridge regression: biased estimation for nonorthogonal problems *Technometrics* **12** 55–67

[37] Hansen P C 1998 *Rank-deficient and Discrete Ill-posed Problems: Numerical Apects of Linear Inversion* (Philadelphia, PA: SIAM)

[38] Krakauer N Y, Schneider T, Randerson J T and Olsen S C 2004 Using generalized cross-validation to select parameters in inversions for regional carbon fluxes *Geophys. Res. Lett.* **31** L19108

[39] Mukherjee A and Zhu J 2011 Reduced rank ridge regression and its kernel extensions *Stat. Anal. Data Min.: ASA Data Sci. J.* **4** 612–22

[40] Chatterjee S, Steinhaeuser K, Banerjee A, Chatterjee S and Ganguly A 2012 *Sparse Group Lasso: Consistency and Climate Applications* (Philadelphia, PA: Society for Industrial and Applied Mathematics) pp 47–58

[41] Lobell D B and Field C B 2011 California perennial crops in a changing climate *Clim. Change* **109** 317–33

[42] McShane B B and Wyner A J 2011 A statistical analysis of multiple temperature proxies: are reconstructions of surface temperatures over the last 1000 years reliable? *Ann. Appl. Stat.* **5** 5–44

[43] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

[44] Biau G 2012 Analysis of a random forests model *J. Mach. Learn. Res.* **13** 1063–95

[45] Francke T, López-Tarazón J A and Schröder B 2008 Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests *Hydrol. Process.* **22** 4892–904

[46] Whitcomb J, Moghaddam M, McDonald K, Kellndorfer J and Podest E 2009 Mapping vegetated wetlands of Alaska using L-band radar satellite imagery *Can. J. Remote Sens.* **35** 54–72

[47] Lima A R, Cannon A J and Hsieh W W 2015 Nonlinear regression in environmental sciences using extreme learning machines: a comparative evaluation *Environ. Modelling Softw.* **73** 175–88

[48] Hoyos I C P, Krakauer N Y and Khanbilvardi R 2016 Estimating the probability of vegetation to be groundwater dependent based on the evaluation of tree models *Environments* **3** 9

[49] Rudiyanto, Minasny B, Setiawan B I, Arif C, Saptomo S K and Chadirin Y 2016 Digital mapping for cost-effective and accurate prediction of the depth and carbon stocks in Indonesian peatlands *Geoderma* **272** 20–31

[50] Wei S, Yi C, Fang W and Hendrey G 2017 A global study of GPP focusing on light-use efficiency in a random forest regression model *Ecosphere* **8** e01724

[51] Krakauer N Y, Lakhankar T and Anadón J D 2017 Mapping and attributing normalized difference vegetation index trends for Nepal *Remote Sens.* **9** 986

[52] Saxe S, Hogue T S and Hay L 2018 Characterization and evaluation of controls on post-fire streamflow response across western US watersheds *Hydrol. Earth Syst. Sci.* **22** 1221–37

[53] Liaw A and Wiener M 2002 Classification and regression by randomForest *R. News* **2** 18–22

[54] Krakauer N Y, Grossberg M D, Gladkova I and Aizenman H 2013 Information content of seasonal forecasts in a changing climate *Adv. Meteorol.* **2013** 480210

[55] Newey W K and West K D 1987 A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix *Econometrica* **55** 703–8

[56] Newey W K and West K D 1994 Automatic lag selection in covariance matrix estimation *Rev. Econ. Stud.* **61** 631–53

[57] Zeileis A 2004 Econometric computing with HC and HAC covariance matrix estimators *J. Stat. Softw.* **11** 1–17

[58] Dorji S, Herath S, Mishra B K and Chophel U 2018 Predicting summer monsoon of Bhutan based on SST and teleconnection indices *Meteorol. Atmos. Phys.* (https://doi.org/10.1007/s00703-018-0589-2)

[59] Pardo-Igúzquiza E 1998 Maximum likelihood estimation of spatial covariance parameters *Math. Geol.* **30** 95–108

[60] Du H and Smith L A 2012 Parameter estimation through ignorance *Phys. Rev.* E **86** 016213

[61] Cai M, Shin C-S, van den Dool H M, Wang W, Saha S and Kumar A 2009 The role of long-term trends in seasonal predictions: implication of global warming in the NCEP CFS *Weather Forecast.* **24** 965–73

[62] Barnston A G and Mason S J 2011 Evaluation of IRI's seasonal climate forecasts for the extreme 15% tails *Weather Forecast.* **26** 545–54

[63] Peng P, Kumar A, Halpert M S and Barnston A G 2012 An analysis of CPC's operational 0.5-month lead seasonal outlooks *Weather Forecast.* **27** 898–917

[64] Aizenman H, Grossberg M D, Krakauer N Y and Gladkova I 2016 Ensemble forecasts: probabilistic seasonal forecasts based on a model ensemble *Climate* **4** 19

[65] Fasullo J 2004 A stratified diagnosis of the Indian monsoon-Eurasian snow cover relationship *J. Clim.* **17** 1110–22

[66] Zhang Y, Li T and Wang B 2004 Decadal change of the spring snow depth over the Tibetan Plateau: the associated circulation and influence on the East Asian summer monsoon *J. Clim.* **17** 2780–93

[67] Shaman J and Tziperman E 2005 The effect of ENSO on Tibetan plateau snow depth: a stationary wave teleconnection mechanism and implications for the south Asian monsoons *J. Clim.* **18** 2067–79

[68] Shaman J, Cane M and Kaplan A 2005 The relationship between Tibetan snow depth, ENSO, river discharge and the monsoons of Bangladesh *Int. J. Remote Sens.* **26** 3735–48

[69] Wu R and Kirtman B P 2007 Observed relationship of spring and summer East Asian rainfall with winter and spring Eurasian snow *J. Clim.* **20** 1285–304

[70] Immerzeel W 2008 Historical trends and future predictions of climate variability in the Brahmaputra basin *Int. J. Climatol.* **28** 243–54

[71] Ravindranath A, Devineni N, Lall U and Concha Larrauri P 2018 Season-ahead forecasting of water storage and irrigation requirements—an application to the southwest monsoon in India *Hydrol. Earth Syst. Sci.* **22** 5125–41
[72] Gillies R R, Wang S-Y, Sun Y and Chung O-Y 2013 Supportive empirical modelling for the forecast of monsoon precipitation in Nepal *Int. J. Climatol.* **33** 3047–54
[73] Wang S-Y and Gillies R R 2013 Influence of the Pacific quasi-decadal oscillation on the monsoon precipitation in Nepal *Clim. Dyn.* **40** 95–107
[74] Nanjundiah R S, Francis P A, Ved M and Gadgil S 2013 Predicting the extremes of indian summer monsoon rainfall with coupled ocean-atmosphere models *Curr. Sci.* **104** 1380–93
[75] Koenker R W and Bassett G W 1978 Regression quantiles *Econometrica* **46** 33–50
[76] Meinshausen N 2006 Quantile regression forests *J. Mach. Learn. Res.* **7** 983–99
[77] Taillardat M, Mestre O, Zamo M and Naveau P 2016 Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics *Mon. Weather Rev.* **144** 2375–93
[78] Bhuiyan M A E, Nikolopoulos E I, Anagnostou E N, Quintana-Seguí P and Barella-Ortiz A 2018 A nonparametric statistical technique for combining global precipitation datasets: development and hydrological evaluation over the Iberian Peninsula *Hydrol. Earth Syst. Sci.* **22** 1371–89