**Severe Precipitation in Northern India in June 2013: Causes, Historical Context, and Changes in Probability**

Deepti Singh, Daniel E. Horton, Michael Tsiang, Matz Haugen, Moetasim Ashfaq, Rui Mei, Deeksha Rastogi, Nathaniel C. Johnson, Allison Charland, Bala Rajaratnam and Noah S. Diffenbaugh\*

**Supplementary Methods:**

***S1: Datasets***

For the precipitation analysis, we use the 1x1 gridded daily dataset from the Indian Meteorological Department (IMD) (Rajeevan et al. 2006, 2010). This dataset is developed from approximately 2140 rain-gauge stations over India, and has been extensively used in literature to study the Indian Monsoon characteristics (Dash et al. 2009; Ghosh et al. 2012; Singh 2013). As shown in Fig. 1a of Rajeevan et al. (2006), all stations included in the development of this dataset have at least 90% data availability within the observational period.

Daily atmospheric variables – including geopotential heights, atmospheric winds, and specific humidity – are obtained from the National Center for Environment Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis (R1) datasets at 2.5x2.5 spatial resolution (Kalnay et al. 1996). In order to match the period of record of the IMD precipitation dataset, we analyze the 1951-2013 period in the reanalysis.

***S2: SOM methodology***

The self-organizing map (SOM) is a neural network-based cluster analysis, similar to k-means clustering, that partitions a high dimensional dataset into a smaller number of representative clusters (Kohonen 2001). In contrast with conventional cluster analysis, these SOM clusters, each of which is associated with a component called a node or neuron, become topologically ordered on a lower-dimensional (typically two-dimensional) lattice so that similar clusters are located close together in the lattice and dissimilar clusters are located farther apart. This topological ordering occurs through the use of a neighborhood function, which acts like a kernel density smoother among a neighborhood of neurons within this low-dimensional lattice. As a result, neighboring neurons within this lattice influence each other to produce smoothly varying clusters that represent the multi-dimensional distribution function of the data used to construct the SOM.

In the present study, we perform two separate SOM analyses to categorize daily June 1951-2013 upper-level (200mb) and lower-level (850mb) geopotential height anomalies in the domain centered over northern India (0-60N, 40-120E). In each SOM analysis, the daily geopotential height anomaly field is treated as an *M*-dimensional vector, where *M* is the number of grid points. The user specifies the number of clusters, *K*, and the final clustering is determined through an iterative procedure that approximately minimizes the Euclidean distance between the daily geopotential height anomaly fields in *M*-dimensional space and their best-matching SOM patterns. During this iterative procedure, the SOM patterns also “self-organize” into a topologically ordered two-dimensional lattice or grid, as described above. Each daily height field is assigned to a best-matching SOM pattern on the basis of minimum Euclidean distance. Additional details of the SOM methodology can be found in the appendix of Johnson et al. (2008). See also Hewitson and Crane (2002) and Liu et al. (2006) for additional discussion of SOM applications in climate science, and Chattopadhyay et al. (2008) and Borah et al. (2013) for SOM applications to Indian monsoon intraseasonal variability. Reusch et al. (2005) and Liu et al. (2006) also compare SOM analysis with empirical orthogonal function analysis for the purpose of pattern extraction, demonstrating several advantages of SOM analysis over the more conventional approach.

As mentioned above, the user must specify the choice of *K* prior to the iterative procedure that determines the SOM clusters. Although there are quite a few approaches for determining an appropriate value of *K*, there is no universally accepted method for determining the optimal *K*. In this study, we make the subjective choice of *K* = 35 (i.e., a SOM organized in a two-dimensional lattice with five rows and seven columns), which is a high enough value to resolve regional pattern detail but small enough to maintain interpretability. We test that our results are robust with different choices for *K*, and that all 35 SOM patterns are statistically distinguishable. For the test of statistical distinguishability, we follow the approach of Johnson (2013), who argues that the maximum number for *K* is the maximum value such that all *K* SOM patterns are statistically distinguishable from each other. This test is administered by evaluating whether or not all *K*(*K*-1)/2 SOM pattern pairs are statistically distinguishable on the basis of a field significance test, which in this case is based on the determination of the “false discovery rate” (Benjamini and Hochberg 1995; Wilks 2006). Additional details of the test are given in Johnson (2013).

In the present application, we add an additional step to account for the strong autocorrelation evident in daily data. Because a SOM pattern typically persists for several consecutive days, the number of daily occurrences of each SOM pattern would substantially overestimate the number of temporal degrees of freedom in the local *t-*tests that determine whether the SOM pattern composite anomalies are significantly different from each other. To correct for this potential source of bias, we perform the local *t*-tests for subsets of the daily geopotential height fields assigned to each SOM pattern. Because the timescale of atmospheric teleconnection patterns is on the order of one to two weeks (e.g., Feldstein 2000), we set the condition that each geopotential height field assigned to the SOM pattern within a subset must be separated by all other geopotential height fields within the subset by at least 15 days. If this separation criterion is not met, then we only keep the daily field with the highest pattern amplitude, where the amplitude is defined as the projection of the daily geopotential height field onto the assigned SOM pattern.

The results of this distinguishability test reveal that all 35 SOM patterns in the 200mb and 850mb geopotential height SOMs are statistically distinguishable from each other at the 5% level. Furthermore, all SOM pattern pairs but one (SOM patterns 2 and 8 in the 850mb geopotential height SOM) are statistically distinguishable from each other at the 1% level. These results suggest that the choice of *K* = 35 is reasonable in that this value of *K* is not too high such that the SOM patterns become statistically indistinguishable from each other. Furthermore, we also have performed the same analysis for *K* = 20 and found that the main features of the 20-pattern SOMs are quite similar to the main features of the 35-pattern SOMs. Therefore, these evaluations support the robustness of the SOM analyses in this study.

From these 35 distinct nodes, we identify the upper- and lower-level atmospheric patterns in June 2013 to track their temporal progression, and assess their historical occurrence and co-occurrence (Fig. 1, Fig. S2). Additionally, we determine typical precipitation patterns associated with each geopotential height SOM pattern (example: Fig. 1i,j, Fig. S2) by creating composites of precipitation from each June day in 1951-2013 on which the atmospheric patterns best matched the corresponding node. This was performed on the basis of assignments of each day to a particular SOM pattern.

***S3: Details of probability quantification***

*S3.1: Determining a parametric fit for the observed distribution*

The event being analyzed lies substantially outside the range of the observed distribution. Therefore, to quantify the probability of occurrence of such an event, we test different parametric distributions (gamma, weibull, lognormal and pareto) to find a suitable fit for the observations (Fig. S3a). We use the maximum log-likelihood estimator to precisely determine the parameters of these potential distributions. Through evaluating the RMS errors (see inset table in Fig. S3a) between the observed and fitted distributions, we conclude that the Pareto-III distribution most closely represents the observed distribution function.

*S3.2: Probability quantification*

We use spatially averaged cumulative June precipitation from 1951-2013 over the flood-affected region (Fig. 1a) to generate an observed distribution to examine the probability of the June 2013 event. After fitting the Pareto-III distribution to the observed cumulative June precipitation, we determine the percentile or probability (p) of a specific event magnitude (Pr), and thereby quantify the return-period (T=1/p) of the event in the context of the observed climate. We find that the total-June 2013 precipitation magnitude exceeds the 99.1th percentile, and has a return period of 111 years (Fig. 2a).

We use the CMIP5 historical (“20C”) and pre-industrial (“PI”) simulations to quantify the influence of observed anthropogenic forcing on the likelihood of the June-2013 total cumulative precipitation. (Fig. 2b-d). The method is illustrated in Fig. S3c. After shifting the “20C” and “PI” distributions by the difference between the observed and 20C means, we use the Kolmogorov-Smirnov (“K-S”) goodness of fit test to measure the closeness of CMIP5-simulated historical (“20C”) distributions to the observed distribution. (Please refer Sperber et al. (2013) for an evaluation of the skill of these models to simulate other characteristics of Indian monsoon rainfall, circulation indices and their teleconnections). Because the simulated change in likelihood of extremes can be heavily influenced by biases in the simulated distribution, we restrict our analyses to 11 models whose K-S value exceeds 0.2 (“A1”; see Fig. S3), ensuring a comparatively good fit of the overall distribution, including in the tails. We then fit the select “A1” CMIP5 models with Pareto-III distributions.

We determine the magnitude of the 99.1th percentile event in the “20C” distribution as representative of the simulated event magnitude (follow vertical red line in Fig. S3c). By definition, the event has a return period of 111 years in the 20C distribution. Next, we find the corresponding percentile \_ and hence return period – of this 20C magnitude event in the preindustrial simulations (follow horizontal blue line in Fig. S3c).

By comparing the likelihood of the event in the pre-industrial climate to the historical climate, the ratio of the return periods (TPI/T20C) provides an estimate of the influence of 20C forcings on the probability of the June 2013 heavy precipitation.

However, the limited sample of precipitation creates uncertainty in the true fit of the Paretto-III distribution to the observed and CMIP5-simulated precipitation data, thereby introducing uncertainty in the PI/20C return period ratio. We therefore use a bootstrapping approach to quantify this uncertainty, and provide confidence estimates for this ratio. We bootstrap the observed, “20C” and “PI” precipitation data to generate many distributions to capture the uncertainty around the true parameters of the Paretto-III fits and thus the uncertainty in the return periods. By applying the probability quantification methodology described above to all these bootstrapped distributions, we generate a distribution of the PI/20C return period ratios, which are represented in box-plots in Fig. 2d.

**References**

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.*, **B57**, 289–300.

Borah, N., A. K. Sahai, et al. (2013). "A self-organizing map–based ensemble forecast system for extended range prediction of active/break cycles of Indian summer monsoon." Journal of Geophysical Research: Atmospheres **118**(16): 9022-9034.

Chattopadhyay, R., A. K. Sahai, et al. (2008). "Objective Identification of Nonlinear Convectively Coupled Phases of Monsoon Intraseasonal Oscillation: Implications for Prediction." Journal of the Atmospheric Sciences **65**(5): 1549-1569.

Dash, S. K., M. A. Kulkarni, U. C. Mohanty, and K. Prasad, 2009: Changes in the characteristics of rain events in India. *J. Geophys. Res*., **114**, D10109, doi:10.1029/2008jd010572.

Feldstein, S. B., 2000: The timescale, power spectra, and climate noise properties of teleconnection patterns*. J. Climate*, **13**, 4430–4440.

Ghosh, S., D. Das, S.-C. Kao, and A. R. Ganguly, 2012: Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes. *Nature Climate Change*, **2**, 86–91, doi:doi:10.1038/nclimate1327.

Hewitson, B.C., and R. G. Crane, 2002: Self-organizing maps: Applications to synoptic climatology. *Climate Res*., **22**, 13–26.

Johnson, N. C., S. B. Feldstein, and B. Tremblay, 2008: The continuum of Northern Hemisphere teleconnection patterns and a description of the NAO Shift with the use of self-organizing maps. *J. Climate*, **21**, 6354–6371.

Johnson, N. C., 2013: How many ENSO flavors can we distinguish?\*. *J. Climate*, **26**, 4816–4827.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc*., **77**, 437–471.

Kohonen, T., 2001: *Self-Organizing Maps*. 3rd ed. Springer, 502 pp.

Liu, Y., R. H. Weisberg, and C. N. K. Mooers, 2006: Performance evaluation of the self-organizing map for feature extraction. *J. Geophys. Res.*, **111**, C05018, doi:10.1029/2011GL047658.

Rajeevan, M., J. Bhate, et al. (2006). "A high resolution daily gridded rainfall for the Indian region : analysis of break and active monsoon spells." Current Science **91**(3): 296-306.

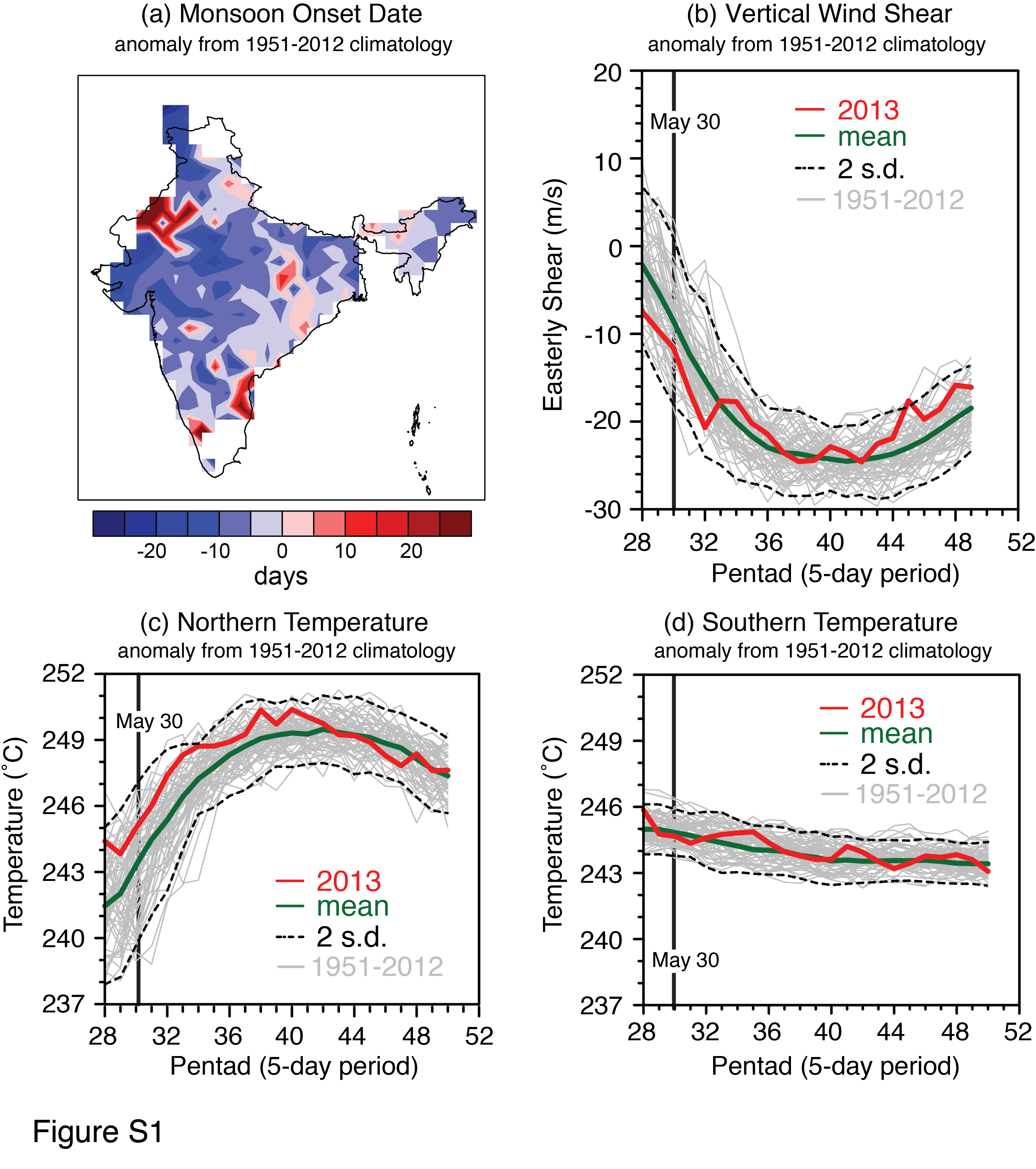
Rajeevan, M., S. Gadgil, and J. Bhate, 2010: Active and break spells of the Indian summer monsoon. *J. Earth Sys. Sci*., **119**, 229–247, doi:10.1007/s12040-010-0019-4.

Reusch, D. B., R. B. Alley, and B. C. Hewitson, 2005: Relative performance of self-organizing maps and principal component analysis in pattern extraction from synthetic climatological data. *Polar Geogr.*, **29**, 188–212.

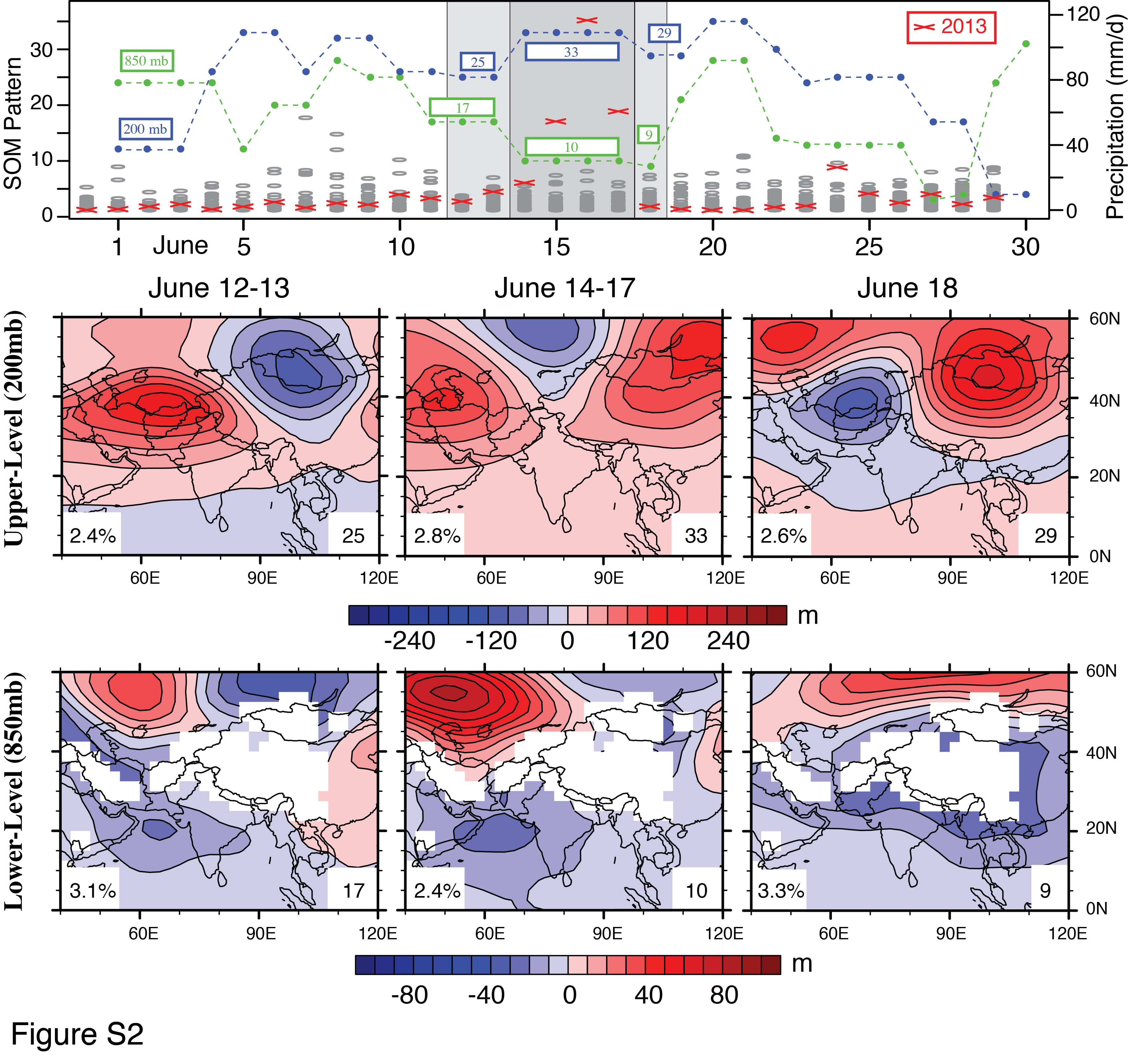
Singh, C., 2013: Characteristics of monsoon breaks and intraseasonal oscillations over central India during the last half century. *Atmos. Res*., **128**,120–128, doi:http://dx.doi.org/10.1016/j.atmosres.2013.03.003.

Sperber, K. R., H. Annamalai, et al. (2013). "The Asian summer monsoon: an intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century." Climate Dynamics **41**(9-10): 2711-2744.

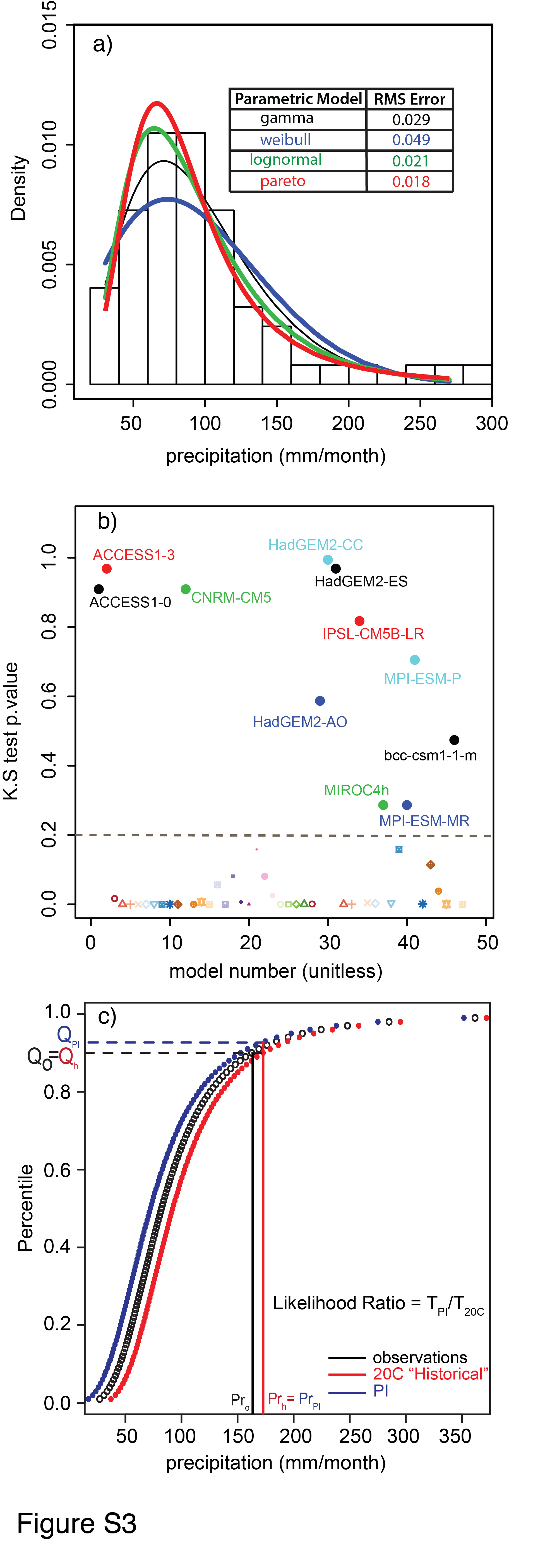
Wilks, D. S., 2006: On “field significance” and the false discovery rate. *J. Appl. Meteor. Climatol.*, **45**, 1181–1189.



**Figure S1 | June 2013 Monsoon Dynamics.** (a) June 2013 monsoon onset anomalies at each grid-cell. The onset date at each grid-cell is defined as the 5-day mean precipitation exceeding the January mean precipitation of that year and 5 mm/day. Negative anomalies over most of central and northwestern India suggest an early monsoon onset in June 2013. (b) June 2013 vertical easterly zonal wind shear. The easterly zonal wind shear is defined as the difference in winds between the upper (200mb) and lower (850mb) atmosphere averaged over the domain (0-30˚N, 50-90˚E). The vertical wind shear was ~2 above the climatological mean prior to the June 2013 event. (c,d) Zonally averaged (52-85˚E) tropospheric temperature (200-500mb) above the land at 30˚N (c) and the ocean at 0˚N (d). The temperature over land (“northern temperature”) also exceeded the climatological mean by 2.



**Figure S2 | Temporal progression of upper- and lower-level synoptic patterns.** We use self-organizing map (SOM) cluster analysis to identify upper- and lower-level atmospheric patterns in June 2013, track their temporal progression, and assess their historical occurrence and co-occurrence (see Supplemental Methods). The SOM algorithm autonomously clusters historical June synoptic patterns into 35 statistically distinct patterns. The time-series in the top panel display the progression of upper-level (blue) and lower-level (green) SOM patterns for each day of June 2013. Patterns relevant to the flooding event are highlighted in grey windows in the time-series panel, and their spatial patterns are displayed in the bottom two panels (see SOM pattern numbers displayed in the lower right of the maps). Daily precipitation accumulation over the flood region (white box Figure 1a) is indicated in the secondary y-axis, with June 2013 values shown in red X’s and each year in 1951-2012 shown in grey ovals. From 11 June to 17 June, blocking patterns persisted in the upper atmosphere (200mb, patterns 25 & 33), while low-pressure troughing strengthened in the lower-atmosphere over the flood region (850mb, patterns 10 & 17). On 18 June, the system lost upper-level support and weakened. The relative frequency of occurrence of each SOM pattern to all historic June days is indicated in the lower left of each SOM map (second and third rows). June 14-17 2013 was the first time upper-level pattern 33 co-occurred with lower-level pattern 10 during the month of June within the 1951-2013 historical record.



**Figure S3 | Statistical framework for probability quantification.** (a) Results from fitting the observed cumulative June precipitation (1951-2012) with 4 parametric models. The Pareto-III model shows the lowest root mean square error between the empirical and model cumulative distribution functions (shown in the table). (b) p-values of the Kolmogorov-Smirnov (“K-S”) goodness of fit test measuring the closeness of CMIP5-simulated historical (“20C”) distributions to the observed distribution. Higher p-values indicate models that more closely simulate the observed distribution, including the tails. (c) Theoretical methodology for quantifying the probability of a specific event in the historical (“20C”) and preindustrial (“PI”) climates. This involves determining the magnitude of the event (Prh) in the “20C” distribution corresponding to the percentile of the observed event (Qo=Qh). We calculate the ratio (PI/20C) of the return periods by estimating the percentile (or probability) of this simulated event (Prh= PrI) in the pre-industrial (TPI) and historical (T20C) distributions.