QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY
Q. J. R. Meteorol. Soc. **00**: 1–17 (0000)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/qj.000

Analysis of Indian monsoon daily rainfall on subseasonal to multidecadal time scales using a hidden Markov model

A. M. Greene[a]* A. W. Robertson[a] and S. Kirshner[b]

[a] *International Research Institute for Climate and Society, USA*

[b] *University of Alberta, Canada*

Abstract:

A 70-year record of daily monsoon-season rainfall at a network of 13 stations in central-western India is analyzed using a 4-state homogeneous hidden Markov model (HMM). The diagnosed states are seen to play distinct roles in the seasonal march of the monsoon, can be associated with "active" and "break" monsoon phases and capture the northward propagation of convective disturbances associated with the intraseasonal oscillation (ISO). Interannual variations in station rainfall are found to be associated with the alternation, from year to year, in the frequency of occurrence of wet and dry states; this mode of variability is well-correlated with both all-India monsoon rainfall and an index characterizing the strength of the El Niño-Southern Oscillation. Analysis of lowpassed time series suggests that variations in state frequency are responsible for the modulation of monsoon rainfall on multidecadal time scales as well.

KEY WORDS Climate diagnosis; Precipitation; Statistical modeling

Received

1 Introduction

Owing to both its meteorological and economic significance, the Indian monsoon has been studied intensively (see, e.g., Gadgil, 2003; Webster et al., 1998; Abrol, 1996; Bharadwaj, 2004; Rai, 2005; Gadgil and Kumar, 2006; Gadgil and Gadgil, 2002). In this contribution, daily monsoon rainfall at a small network of stations is decomposed using a hidden Markov model (HMM). Although such models are themselves no longer novel, to the best of our knowledge they have not heretofore been deployed in this regional setting; investigation of their capabilities in the monsoon domain is therefore of interest. The HMM is utilized here as a diagnostic tool. However, its potential applicability extends as well to the downscaling of climate forecasts and the simulation of weather-within-climate, in the context of climate change. The present analysis may thus be considered an evaluative first step in the utilization of such models.

*Correspondence to: International Research Institute for Climate and Society, Palisades, NY 10964 USA. E-mail: amg@iri.columbia.edu

The HMM associates observed patterns of daily rainfall with a small set of "hidden states," which proceed in time as a first-order Markov process (Hughes and Guttorp, 1994; Hughes et al., 1999). It may be considered a parsimonious description of the raw rainfall observations, and provides a simple means of generating synthetic precipitation series having the same statistical properties (including spatial distribution) as the data to which it is fit.

Monsoon rainfall is highly variable both temporally and spatially, in particular at the scale of individual weather stations. The HMM is fit directly to the daily station data, sans any filtering or gridding, yet is shown to capture functional aspects of rainfall variability across a broad range of time scales. This suggests the existence of some scale-bridging mechanism. It has been noted, e.g., that the "active" and "break" phases that characterize subseasonal monsoon variability are also expressed, via aggregation, in terms of interannual variations (Gadgil and Asha, 1992; Gadgil, 1995; Goswami and Xavier, 2003; Goswami, 2005). At the other end of the spectrum, it is of interest to see whether such aggregation might play a role in decadal fluctuations as well.

The link between the monsoon and El-Niño-Southern Oscillation (ENSO) has also received attention, (Rasmusson and Carpenter, 1983; Shukla, 1987; Krishnamurthy and Goswami, 2000; Kumar et al., 2006), the general consensus being that strong El Niño events tended to be associated with weak monsoons, at least up until the late 1970s (Kumar et al., 1999). This linkage is explored here through examination of interannual variations in the frequency of occurrence of the diagnosed states and the NINO3.4 index (Barnston et al., 1994).

Section 2 describes the datasets employed and Sec. 3 some climatological characteristics of the data. The HMM is discussed in Sec. 4, while Sec. 5 examines the hidden states in terms of associated atmospheric composites. Sections 6, 7 and 8 deal with intraseasonal, interannual and multidecadal variability, respectively; a discussion and summary follow, in Secs. 9 and 10.

2 Data

A 70-yr record of daily rainfall data at 13 stations, from the Global Daily Climatology Network (GDCN, Legates and Willmott, 1990), constitutes the primary dataset. These stations were initially chosen to match, by name and location, a group of records for 1973-2004 that had been obtained from the Global Surface Summary of Day Data (NCDC, 2002), with the thought that the two datasets could be combined. However, given the reasonably long GDCN data series and their relative freedom from missing values, as well as potential inhomogeneity issues, we restrict our attention to this 70-yr record.

The GDCN data span the years 1901-1970, with a station average of only 11 missing days out of 8540, and no station missing more than 29 days. The stations themselves are listed in Table I and locations shown in Fig. 1. Although some regions, notably the eastern coast, are not sampled, atmospheric circulation composites (discussed in Sec. 5) suggest that this network captures enough of the spatiotemporal variability of the precipitation field for the large-scale features of the monsoonal circulation to be quite well-inferred. Interannual variations in mean station rainfall are well-correlated ($r = 0.86$) with the Indian Summer Monsoon Rainfall (ISMR) index, an average of JJAS rainfall over approximately 300 stations (Sontakke et al., 1993). Rainfall occurrence probability and mean daily intensity (amount, conditional on the occurrence of rain) for stations in the network agree well with values derived from a high-resolution (1°) gridded daily dataset

DOI: 10.1002/qj

Table I. Stations whose records are utilized herein. Columns, from left, indicate station number (corresponding to the numbers shown in Fig. 1a), station ID as provided in the dataset, station name, latitude, longitude, climatological rainfall probability and mean rainfall amount on wet days.

No.	ISTA	Name	Lat (N)	Lon (E)	P(R)	$\overline{I}$
1	5010600	Ahmadabad	23.06	72.63	0.41	15.0
2	5100100	Veraval	20.90	70.36	0.43	11.0
3	5150100	Rajkot	22.30	70.78	0.35	13.7
4	5171200	Surat	21.20	72.83	0.56	15.3
5	11170400	Indore	22.71	75.80	0.54	13.1
6	11180800	Jabalpur	23.20	79.95	0.58	17.4
7	12040300	Aurangabad	19.85	75.40	0.49	10.2
8	12190100	Poona	18.53	73.85	0.58	7.2
9	12230300	Sholapur	17.66	75.90	0.41	10.9
10	19070100	Bikaner	28.00	73.30	0.17	12.1
11	19131300	Jaipur	26.81	75.80	0.36	12.1
12	22021900	Delhi	28.58	77.20	0.30	16.0
13	23351200	Lucknow	26.75	80.88	0.43	17.1

from the India Meteorological Department, (IMD, Rajeevan et al., 2006).

Atmospheric circulation fields are derived from the National Centers for Environmental Prediction-National Center for Atmospheric Research (NCEP-NCAR) reanalysis (Kalnay et al., 1996). Comparisons were made between composites derived from this dataset and from the European Centre for Medium-Range Weather Forecasts ERA-40 reanalysis (Uppala, 2001). Those from the latter product were found to be somewhat noisier, perhaps owing to the shorter usable data length (period of overlap with the rainfall data, which is about half as long for ERA-40). The NCEP-NCAR data were therefore utilized.

3 Climatology

3.1 Rainfall — spatial distribution

Figures 1a and 1b illustrate mean Jun-Sep (JJAS) climatological rainfall probabilities and mean daily intensities, respectively, over the network. Topographic contours from the GLOBE digital elevation model (Hastings and Dunbar, 1998) are also shown, and the stations and corresponding data are listed in Table I. Both probabilities and intensities are computed conditional on a minimum daily rainfall amount of 0.1 mm, the minimum non-zero value recorded in the dataset.

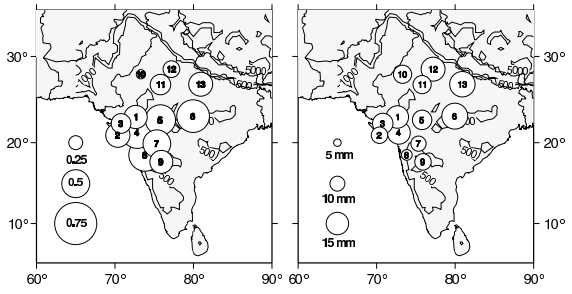

Figure 1. Climatological (a) occurrence probability and (b) mean daily intensity (mm) for Jun-Sep (JJAS) 1901-1970.

In Fig. 1, high probability and intensity values along the western coast reflect onshore flow, as shown in Fig. 4a, impinging on coastal orography (Gadgil, 2003). This orography can also produce sharp gradients in intensity, as is the case with station 8. The high values at station 6, on the other hand, arise from convective systems propagating northwestward from the Bay of Bengal, reaching across the classical "Monsoon zone," a broad belt extending across the midsection of the subcontinent (see Fig. 5a in Gadgil, 2003). The values shown on Figs. 1a and 1b are generally consistent with the IMD 1° data, including the low probabilities and intensities to the north and northwest, such as those of station 10, Bikaner, in arid Rajasthan, as well as the N-S intensity gradient along the southwestern coast.

3.2 Rainfall — seasonal cycle

Figures 2a and 2b shows climatological seasonal cycles for mean occurrence probability and mean daily intensity for each of the 13 stations, in terms of pentads. Both variables exhibit a decided seasonal cycle at all stations. Station 10 again stands out for its relatively low occurrence probabilities, although it clearly participates in the seasonal cycle. It is less of an outlier in terms of mean intensity; this can be seen as well on the maps of Fig. 1.

DOI: 10.1002/qj

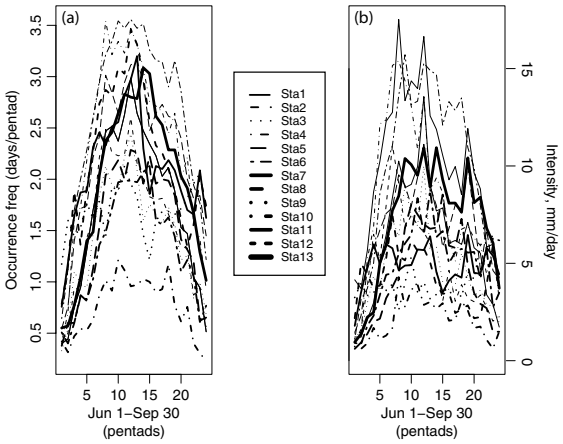

Figure 2. Climatological seasonal cycle of mean (a) occurrence probability and (b) mean daily intensity (mm) for the 13-station dataset, 1901-1970. Data is conditional on a daily threshold of 0.1 mm.

3.3 Circulation

Figure 4a shows mean climatological JJAS horizontal winds at 850 mb and the 500-mb vertical velocity (ω) from the NCEP-NCAR reanalysis, for 1951-1970. Several well-known features of the monsoon circulation are evident in the plot, including the cross-equatorial Somali jet along the coast of Africa, the southwesterly to westerly flow in the Arabian sea, and ascending motion ($\omega < 0$) concentrated in maxima in the eastern Arabian sea and central Bay of Bengal (Goswami, 2005; Xie and Arkin, 1996).

4 Model description

4.1 Basic structure and assumptions

The HMM factorizes the joint distribution of historical daily rainfall amounts recorded on a network of stations in terms of a few discrete states, by making two assumptions of conditionality: First, that the rainfall on a given day depends only on the state active on that day, and second, that the state active on a given day depends only on the previous day's state. The latter assumption corresponds to the Markov property; the states are considered "hidden" because they are not directly observable.

The state-associated rainfall patterns comprise a probability distribution function for daily rainfall for each of the stations. In the present instance these are three-component mixtures, consisting of a delta function to represent zero-rainfall days and two exponentials representing rainfall intensity. Mixed-exponential distributions have been found effective in the representation of daily precipitation (Woolhiser and Roldán, 1982). Daily rainfall, conditional on state, is represented as

$$p(R_t^m = r | S_t = i) = \begin{cases} p_{im0} & r = 0 \\ \sum_{c=1}^{C} p_{imc} \lambda_{imc} e^{-\lambda_{imc} r} & r > 0 \end{cases} \quad (1)$$

where indices i, m and c refer to state, station and mixture component, respectively and the p_{imc} are weights. In the summation, $C = 2$, i.e., two exponentials are utilized. We do not dwell further on details of the model here; for a more complete description see Robertson et al. (2004, 2006). Note that the Markov transition matrix, together with the rainfall probability density functions, are sufficient for the generation of stochastic rainfall sequences.

4.2 Model selection and fitting

The number of states to be modeled must be specified a priori; differing modeling objectives may lead to differing choices in this regard. A small number of states facilitates diagnosis and model comprehensibility, while a larger number might enable a technically better fit, the latter being more suitable for the generation of synthetic data. For this report, models having three, four, five and eight states were examined in detail. Of these, the three-state model was determined to be suboptimal, particularly in relating the retrieved states to the propagating convective disturbances characteristic of the intraseasonal oscillation

DOI: 10.1002/qj

(ISO, see Sec. 6.3). On the other hand, the five-state model does not add anything new to the descriptions already present with four states and begins to exhibit "state splitting," the subdivision of attributes among states. Examination of the eight-state model confirms the tendency for complexity, but not necessarily clarity, to increase with the number of states. The Bayesian Information Criterion (Schwarz, 1978) was applied to fitted models having up to 20 states, and indicated that overfitting did not occur when as many as 10 states were modeled. For the illustrative purposes that are primary here, a model having four hidden states was thus selected. We show in the following sections that this model provides a physically meaningful description of the monsoon and its variability across a wide range of time scales.

Parameter estimation is carried out by maximum likelihood, using the iterative expectation-maximization (EM) algorithm (Dempster et al., 1977; Ghahramani, 2001). The algorithm was initialized ten times from random starting points; the run utilized was that having the highest log-likelihood. Estimation was performed using the Multivariate Nonhomogeneous Hidden Markov Model (MVNHMM) Toolbox, developed by one of the authors (Kirshner, see http://www.cs.ualberta.ca/˜sergey/MVNHMM/).

4.3 Representation in terms of states

Fig. 3 shows rainfall occurrence probabilities and mean intensities for the 4-state model, the former derived from the p_{im0} parameter in Eq. 1, the latter the means of the mixed-exponential distributions. The four states exhibit distinctly different patterns for both variables: State 4, which might be characterized as the "dry" state, exhibits small occurrence probabilities, and mean intensities that are small to moderate, at all stations, while state 3, the "wet" state, exhibits relatively high occurrence probabilities and intensities. State 1 is also rather wet but shows a substantial NE-SW gradient in mean intensity, while state 2 exhibits a S-N gradient in both occurrence probability and mean intensity.

4.4 Transition matrix

From a mechanistic (or generative) point of view, the HMM transition matrix provides the stochastic "engine" that drives the system from state to state with the progression of days. Alternatively, the matrix can be regarded as descriptive, summarizing the temporal dependence of the observations in probabilistic form. The entry in row i, column j gives the conditional probability of an i-j transition, i.e., the probability that tomorrow's state will be j, given that today's is i. The matrix for the 4-state HMM is given in Table II. Note that the 3-4 and 4-3 probabilities are both zero (to at least seven decimal places); in fact, no direct transitions between these two states occur. This suggests that some *dynamical* process, that mediates transitions between the intense endmember conditions represented by these two states, has been encoded in the state descriptions. This would be consistent with the conjecture of Ghil and Robertson (2002), that the states represent "slow phases" of a cyclic behavior, here presumably the ISO. This is discussed further in Sec. 6.3.

The largest values in the transition matrix lie along the leading diagonal. Since these are the "self-transition" probabilities (i.e., probabilities of remaining in the respective states), this signals a tendency, for all the states, to persist beyond the length of the sampling interval (one day). It is this tendency that accounts for the horizontally banded appearance of Fig. 5a, the most-likely state sequence (see Sec. 6.1).

The values shown in Table II are *time-invariant.* As will be shown, however, there is a pronounced seasonal

DOI: 10.1002/qj

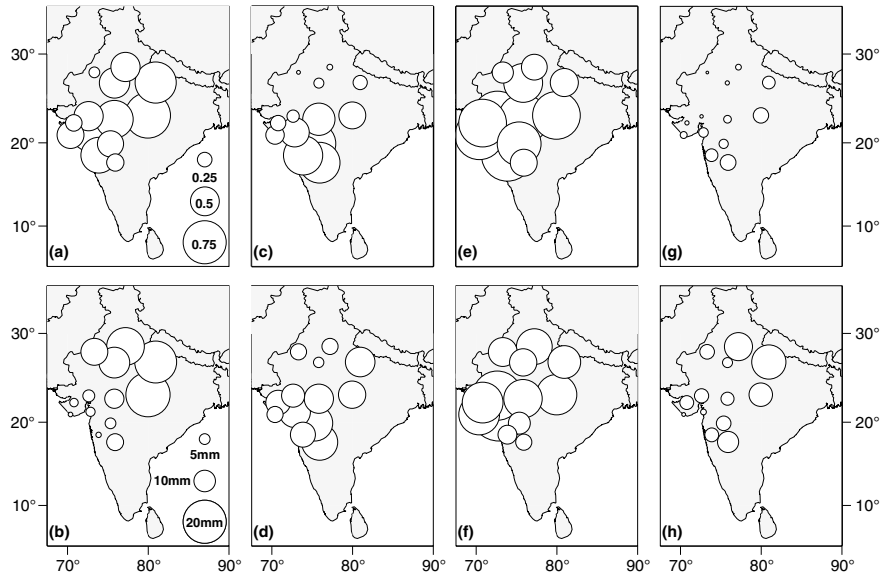

Figure 3. Occurrence probabilities (top) and mean daily intensities for the 4-state model, states 1-4 shown from left to right. Values are computed conditional on mean daily intensity of at least 0.1 mm. Legend in leftmost figures applies to all plots in the respective rows.

Table II. Transition matrix for the 4-state HMM. "From" states occupy the rows, "to" states the columns. Thus, e.g., the probability of a transition from state 2 to state 4 is 0.056.

		"To" state			
		1	2	3	4
"From" state	1	0.798	0.031	0.087	0.083
	2	0.059	0.776	0.109	0.056
	3	0.160	0.018	0.823	0.000
	4	0.022	0.101	0.000	0.876

cycle in the relative frequency of occurrence of the various states, implying that in actuality the transition probabilities must be temporally modulated. In order to drive a fully time-dependent HMM, some coupling to the time dependence of the large-scale fields must evidently be modeled. The corresponding statistical structure is referred to as a "non-homogeneous" HMM (Hughes and Guttorp, 1994).

5 Atmospheric correlates

It is of interest to see how the modeled states are related to the large-scale circulation, since the latter provides a primary control on rainfall. Figures 4b-4e show composited anomalies (with respect to the climatology shown in Fig. 4a) for states 1-4, respectively, for the 4-state HMM. These are constructed by first estimating the most-likely state sequence (see Sec. 6), then averaging the wind fields over those days attributable to each of the states.

State 3 (Fig. 4d), the wet state, is seen to be associated with an *amplification* of the mean seasonal pattern, along with landward displacements of the two centers of ascending motion that straddle the subcontinent. Anomalous cyclonic circulation is present over these centers. There is also anomalous onshore flow from the Arabian sea, as well as an anticyclonic circulation to the southwest of the Indian peninsula. Such departures are consistent with the high occurrence probabilities and intensities on the western coast, extending toward the northeast, that characterize this state (Figs. 3e and 3f). In addition, the more-or-less zonal band of anomalous ascending motion can be identified with the often-described "monsoon trough," (see, e.g., Rao, 1976). Here, this band extends broadly in a northwestward direction from the Bay of Bengal, lying squarely over the monsoon zone as identified by Gadgil (2003). This atmospheric configuration also corresponds

DOI: 10.1002/qj

well with the dominant mode of intraseasonal variability identified by Annamalai et al. (1999) via empirical orthogonal function (EOF) analysis, and identified with the "active" monsoon phase (Goswami, 2005).

The composite for state 4 (Fig. 4e), the "dry" state, is almost a mirror image of that of state 3, with anomalous *descent* over much of the subcontinent, and anticyclonic circulation anomalies where state 3 presents cyclonic ones. This pattern is consistent with the lower occurrence probabilities and intensities shown in Fig. 3g and 3h.

State 1 exhibits higher occurrence probabilities than state 3 for some stations in the northeastern part of the domain; the band of anomalous ascending motion seen in the corresponding composite (Fig. 4b), although configured somewhat differently than in state 3, can still be seen to correspond to the classical monsoon trough. However, the mean intensities here show a strong *gradient* (Fig. 3b), with values increasing from southwest to northeast. In the composite we see anomalous anticyclonic circulation centrally located in the Arabian sea, while a single center of anomalous ascent occurs near the northeastern part of the domain, near the Himalayan escarpment. Thus, high pressure brings dry conditions to the coastal stations, while the anomalous ascent is located in a position consistent with higher inland rainfall intensities.

Finally, state 2 (Fig. 4c) shows anomalous easterly inflow from the Bay of Bengal, resembling in this sense the dry state. State 2 also mirrors state 1 (cf. Fig. 4b), in that there is now anomalous cyclonic activity, with associated rising motion, near the southwest coast, while the north-central region shows anomalous descent and anticyclonic circulation. These properties, like those of the other composites, are reflected in the maps of occurrence probability and mean intensity (Figs. 3c and 3d).

Broadly speaking, the circulation anomalies associated with states 1 and 2 appear as weak versions of those associated with 3 and 4, respectively, but with centers of action displaced to the north. States 1 and 3 may both be considered "wet," and show anomalous ascent over the monsoon zone, while state 2 shows affinities to the dry state, most clearly in the northern part of the domain. Among the questions that then naturally arise is that of the relationship, if any, between the diagnosed states and the wet and dry intraseasonal-scale phenomena known as "active" and "break" phases of the monsoon cycle. This question is taken up in Sec. 6, along with some other aspects of intraseasonal variability as viewed through the prism of the state decomposition. In any event, the patterns of rainfall associated with the diagnosed states appear to correspond quite sensibly with large-scale monsoon-related circulation regimes (Gadgil, 2003).

6 Intraseasonal variability: Breaks and ISO

6.1 Viterbi sequence

Once the parameters of the HMM have been estimated, the most-likely daily sequence of states can be determined using the Viterbi algorithm (Forney, Jr., 1978), a dynamic programming scheme. The Viterbi sequence, which expresses the time evolution of rainfall patterns over the entire data period in terms of the hidden states, is shown for the 4-state model in Fig. 5a. Figure 5b shows the climatological sequence for 1901-1970 (i.e., averaging over the 70 years).

Figure 5a reveals a systematic progression in state occurrence over the course of the season, in the presence of considerable variability on interannual, as well as longer, time scales. During the first half of June, state 4, the dry state, predominates, while during the core of the rainy season states 1 and 3 take center stage. State

DOI: 10.1002/qj

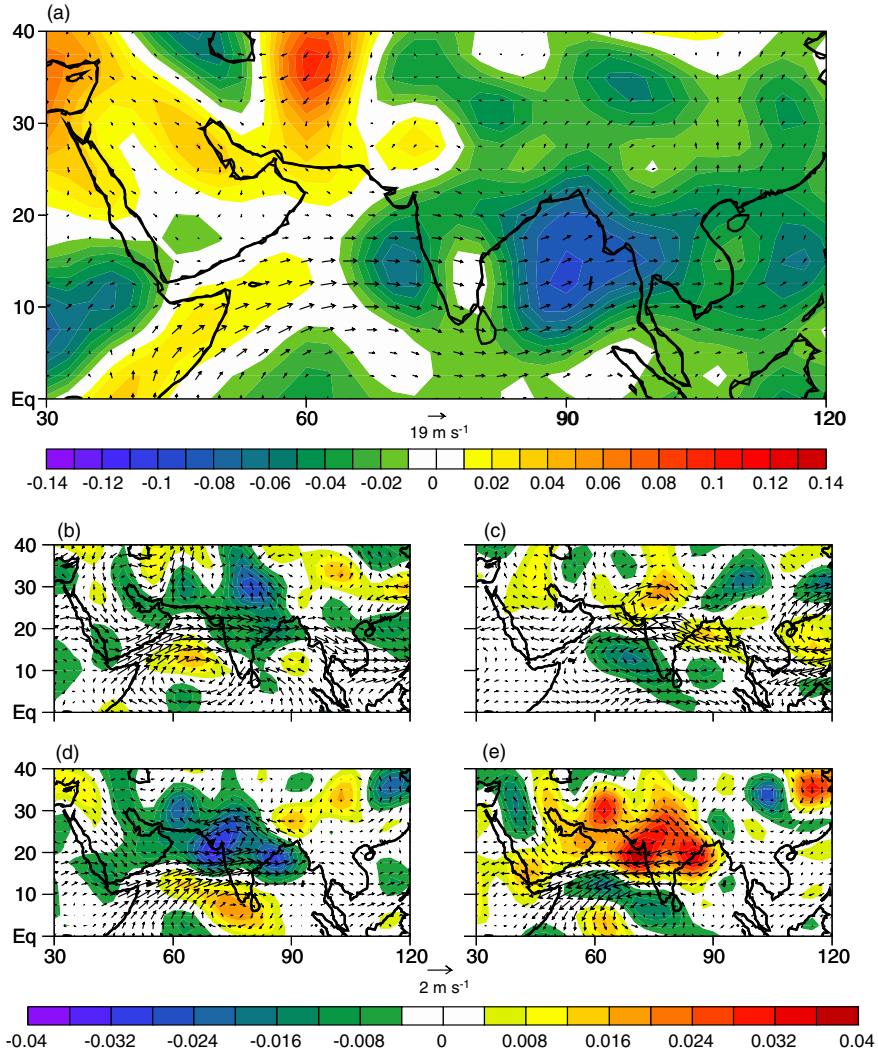

Figure 4. (a) Horizontal winds at 850 mb (vectors, m s^{-1}) and 500-mb vertical velocity (colors, Pa s^{-1}) for the mean climatology. (b)-(d) Same as (a), but for states 1-4, respectively.

2 appears in a quasi-transitional role, first appearing as a bridge between dry and wet conditions in late June, then returning for a brief encore near the end of the season. The dry state again assumes dominance after mid-September. Figure 5b shows in addition that during the central wet period, predominance tends to shift from state 3 in July, toward state 1 in August. Over the 70-year data period the four states occur on an average of 34, 22, 30 and 36 days, respectively, during the 122-day (JJAS) season, with standard deviations 10.3, 7.9, 11.8 and 15.2 days. Thus, considerable interannual variability is present.

6.2 Monsoon breaks

Figures 5a and 5b both show clearly the dominance of state 4 during the early and late stages of the monsoon. This state also occurs sporadically during the July-August (JA) core of the rainy season, however, suggesting a possible association with monsoon breaks. Gadgil and Joseph (2003) provide a listing of breaks for 1901-1989, as defined by rainfall thresholds in the western and eastern sectors of the monsoon zone. These thresholds were chosen in order that there be a good correspondence

Prepared using qjrms3.cls
DOI: 10.1002/qj

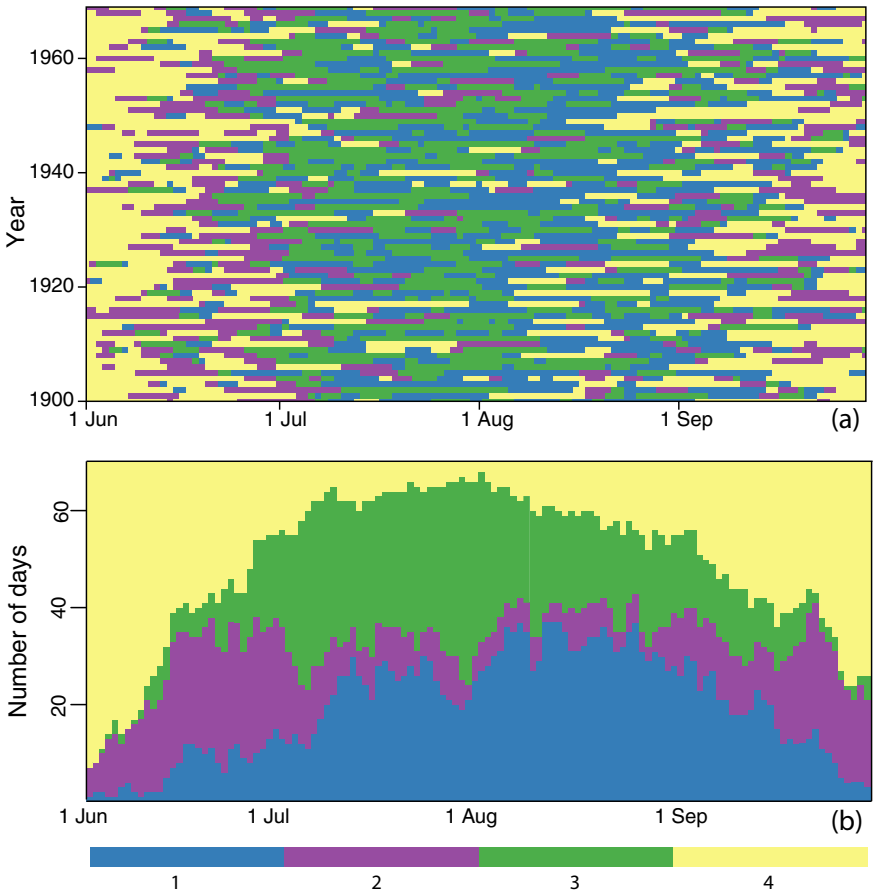

Figure 5. (a) Viterbi sequence of most likely states for the base model, and (b) the corresponding seasonal climatology. Colorbar indicates states 1-4.

between breaks so defined and breaks as identified in a broad range of other studies, so their listing can be considered representative.

Gadgil and Joseph identify an average of 8.8 break days during JA for 1901-1970, while the average number of dry-state days is 7.9. Standard deviations of the Gadgil and Joseph and state 4 series are 6.3 and 7.4 days, respectively. There are nine years in which the Gadgil and Joseph listing shows no break days, and in each of these years there are no occurrences of the dry state. However, there are five additional years in which the dry state does not occur, in which Gadgil and Joseph do indicate breaks. For these years the average number of break days listed is 7.6, only slightly below the mean value. Interannual variations in the number of break and state-4 days are highly correlated ($r = 0.76$, significant at the 0.0001 level in a two-sided test).

Correspondence between the *particular* days when breaks are diagnosed and those days when the Viterbi algorithm identifies state 4 can be expressed in the form of a 2×2 contingency table and evaluated by means of the χ^2 test, either summing over years, or considering the entire dataset as a single long sequence. In either case, the number of overlapping days (332 for the 70-year sequence) significantly exceeds the number expected by chance alone (78), the test statistic being significant at values beyond software precision.

To investigate whether serial autocorrelation might

DOI: 10.1002/qj

bias the expected number of overlap days, the data were resampled, computing the number of such days between differing years chosen at random from the two datasets. The expected number of overlap days per season computed from 2000 samples, was 1.14, whereas the corresponding value calculated directly from the mean occurrence probabilities is 1.12, a difference too small to affect the significance tests. We conclude that serial autocorrelation has not biased the results of these tests, and thus that there exists a high degree of correspondence between occurrence of the HMM dry state and monsoon break days as diagnosed by Gadgil and Joseph.

One characteristic break pattern defined by the IMD is described thus: "There are periods when the monsoon trough is located close to the foothills of the Himalayas, which leads to a striking decrease of rainfall over most of the country, but increase along the Himalayas, parts of northeast India and southern peninsula." (cited in Gadgil and Joseph, 2003). With respect to the composites, this situation would appear to correspond most closely to state 1, which shows a region of anomalous ascent located near the Himalayan foothills (Fig. 4b). Rainfall occurrence probabilities for state 1, however, are uniformly moderate to high (Fig. 3a), quite different from those of the dry state.

Clues to this conundrum may be found in the amount distribution, shown in Fig. 3b, and, somewhat more cryptically, in the Viterbi sequence (Fig. 5a). In the former figure, amounts in the southern part of the domain are seen to be small compared to those in the north, consistent with northward migration of the zone of intense rainfall, while close examination of the latter reveals that nearly all occurrences of the dry state are proceeded by state 1, despite the comparable prevalence of state 3 during JA (1693 and 1627 days for states 1 and 3, respectively). Given these frequencies, it is striking that of the 88 state-4-diagnosed breaks, 81 are immediately preceded by state 1, the remainder by state 2.

These observations suggest that state 1 may describe a *phase* in the northward propagation of monsoon disturbances (Annamalai et al., 1999), occurring as a low-pressure trough reaches the Himalayan foothills but before the anomalous large-scale sinking motion associated with state 4 has become established. This notion is corroborated by the null probability of 3-4 and 4-3 state transitions, as discussed in Sec. 4.4.

6.3 Intraseasonal oscillation

Within-season monsoon variability has been described in terms of the so-called intraseasonal oscillation (Annamalai et al., 1999; Goswami and Mohan, 2001; Goswami, 2005), a quasi-cyclical behavior having a rather broad spectral signature, but with principal activity in the 10-20 and 30-60 day bands (Goswami and Mohan, 2001). Two centers of convective activity are involved, one extending along the monsoon trough, which is then characterized as a tropical convergence zone (TCZ) and extends from the northern Bay of Bengal northwestward over the Indian landmass, and a second lying in the Indian Ocean between 0° and 10° S. The detailed time evolution of the ISO is apparently complex, consisting, according to Goswami and Mohan (2001) of "... fluctuations of the TCZ between the two locations and repeated propagation from the southern to the northern position..." Annamalai et al. (1999) in fact refer to the northward propagation of convective activity as "nonperiodic." In any event, the two "phases" of the ISO, i.e., with convective centers of action located over the two preferred zones, are to be associated with the active and break phases of the monsoon, the northerly location corresponding to the active phase.

DOI: 10.1002/qj

In light of this description, states 1 and 3 (Figs. 4b and 4d) can clearly be identified with the active phase, while state 4, and to a lesser extent, state 2, may be identified with the break phase. However, for the latter two states there is little in the vertical motion field south of the equator (region not shown in these plots) to suggest deep convection. Thus, while some aspects of a correspondence between the state composites and the ISO seem reasonably clear, the structure of the dry state does not appear to correspond in all particulars to the canonical break-phase description of Goswami and Mohan (2001).

The HMM is sensitive not only to differing patterns of rainfall occurrence and intensity per se, but also to the number of days on which these patterns are manifest. Thus, a distinctive pattern that occurred on only a very small number of days would tend to be subsumed into a state having greater representation among the observations. A propagating pattern would then most likely find expression in terms of its more temporally persistent phases. Ghil and Robertson (2002) consider the relationship between persistence, atmospheric states and oscillatory modes in the context of a "wave-particle duality," the modes, or "slow phases," in their terminology, thus more likely to be captured by the state descriptions.

Figure 6 shows composites of 850-mb relative vorticity corresponding to the wind fields of Figs. 4b-4e. Transition probabilities for the JA core of the wet season, estimated from the Viterbi sequence, are shown in Table III (cf. Table II, which applies to the entire JJAS season). The *off-diagonal* elements in this array indicate that a most-likely sequence for the evolution of states would be 4-2-3-1, assuming we begin with a break, and simply 2-3-1-2-3-1... if we do not. The vorticity composites, viewed in either of these sequences, are consistent with a northward-propagating wave. Mean spell lengths from the Viterbi sequence for states 1-4 during JA are 3, 13, 8 and 3 days, respectively, suggesting average cycle lengths of 27 and 24 days for the sequences begging with state 4 and state 2, respectively. These periods fall between the 10-20 and 30-60 day periods discussed by Gadgil and Asha (1992) or Goswami and Mohan (2001) and may thus represent mixtures of the two modes.

Table III. JA transition probabilities (cf Table II).

		"To" state			
		1	2	3	4
"From" state	1	0.847	0.014	0.091	0.047
	2	0.066	0.763	0.158	0.013
	3	0.132	0.009	0.859	0.000
	4	0.050	0.106	0.000	0.844

From Table III, and in regard to the sequence given above, it can be seen that the 1-3 transition probability is about twice that of 1-4. An alternation between states 1 and 3 is consistent with the maintenance of generally heavy precipitation during JA, while the excursions to state 4 comport with the occasional occurrence of breaks. Stochastic switching between these two transitional modes would be consistent with the intermittent character of northward propagation associated with the ISO, as described by both Annamalai et al. (1999) and Goswami (2005).

One other feature of interest in Fig. 5b involves the shift in dominance, during the peak JA period, from state 3 toward state 1. This may reflect an increasing tendency toward the dry state (nearly always preceded by state 1 but never by state 3), and ultimately the end of the rainy season itself, as July turns to August. Increasing predominance of state 1 as the season matures may also be viewed as a tendency, with time, for convection to occur preferentially in the more northerly reaches of the country.

DOI: 10.1002/qj

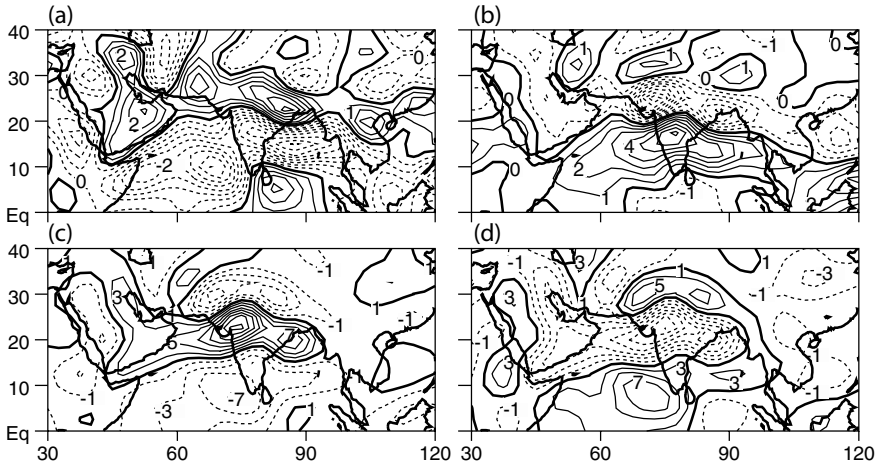

Figure 6. Relative vorticity anomaly composites for the 4-state model, states 1-4 shown in panels (a)-(d), respectively. Units are 10^{-6} s^{-1}.

In summary, much in the state composites is consistent with the ISO, as it has been variously described. However, it should be remembered that the states are not regular snapshots in time, constrained to follow one another in a deterministic order. Furthermore, the data contain information about all time scales; they have not been filtered to retain only ISO-band variability.

7 Interannual variations — Influence of ENSO

The number of days in a given year assigned to each of the states may be computed from the Viterbi sequence. Correlation coefficients for the four frequency-of-occurrence (FO) series thus obtained and the NINO3.4 index are -0.18, -0.16, -0.45 and 0.56 for states 1-4, respectively. The first two of these values are not statistically significant (two-sided test), even at a level of 0.10, while the latter two do prove significant at better than 0.001 (on 68 d.o.f.) This indicates a tendency for El Niño (La Niña) years to be associated with increased FO of the dry (wet) state, consistent with the sense of the historical ENSO-monsoon relationship. The NINO3.4 index is also anticorrelated with the ISMR ($r = -0.63$), indirectly linking FO to this broad-scale metric.

The correlation values provided above do not constitute a complete description of the relationship between FO and station rainfall, in part because FO need not (indeed, cannot) vary independently among states. In addition, *within-state* variation might act to either amplify or compensate year-to-year changes in FO. We thus turn to canonical correlation analysis (CCA, see, e.g., Wilks, 2006) in order to characterize the FO-rainfall relationship. CCA identifies pairs of patterns across two fields, such that the temporal correlation between members of a pair is maximized. The original variables can be projected onto the diagnosed patterns to estimate the degree to which the actual behavior of the fields is captured by them. In the CCA performed here, the method of Barnett and Preisendorfer (1987), in which the original data are first expressed, or "filtered," in terms of EOFs, is utilized. Moron et al. (2006) have performed a similar analysis in an investigation of Senegalese rainfall.

The two "fields" analyzed, each having annual values, are the state FO series and the station seasonal rainfall amounts. Initially, all series are filtered to remove decadal and longer-period variability. This is done by first generating smoothed versions of the series, using 11-yr running

DOI: 10.1002/qj

means. These smoothed versions are then subtracted from the original series, leaving the shorter-period variations. The Kolmogorov-Smirnov test does not lead to a rejection of the null hypothesis of normality for any of the resulting state or station series; the CCA was thus applied directly to these fields, sans any transformation of variables.

Figures 7a and 7b illustrate, respectively, the FO and station rainfall patterns corresponding to the leading mode of covariability. The correlation between the two canonical variates for this mode is 0.92, while the patterns themselves explain 48% of the variance of the FO field and 33% of the variance of the rainfall amounts. A Monte Carlo significance test that involves scrambling the time indices while retaining spatial field structure indicates that the correlation value is significant at better than 0.001. The next two modes also have significant correlation coefficients and explain 14% and 12% of the rainfall variance, respectively. Thus, the leading coupled mode on subdecadal time scales consists of an *alternation* between states 3 and 4, the wet and dry states, coupled to a rainfall pattern in which mean seasonal amounts change in the same sense at all stations, becoming wetter (drier) when state 3 (4) predominates. Thus, from the HMM perspective, ENSO modulates monsoon rainfall through the agency of the state frequencies, producing lower (higher) counts for state 3 (4) in El Niño years, and vice versa for La Niña years.

The leading canonical variate time series for the FO series is well-correlated with the ISMR index ($r = 0.81$, significant at better than 0.0001). This can be taken as additional confirmation that the HMM state decomposition, based on only a 13-station network, has captured patterns that are implicitly descriptive of this broadly representative index.

Utility of the HMM as a predictive downscaling tool was tested for the interannual case by attempting to forecast precipitation over the station network for each year, using a CCA fitted to the remaining data years. The four FO series were utilized as predictors, and all three significantly correlated CCA modes, which together explain 60% of the station rainfall variance, were utilized. The correlation between observed and cross-validated forecast station rainfall series was 0.49 ± 0.13 (1σ), and the mean RMS error 1.7 mm, or 30% of the seasonal mean daily amount (averaged over both stations and years). For the stations with higher correlations this represents potentially useful forecast skill. It should be kept in mind, however, that these measures assume a perfect forecast of the state frequencies, and thus describe only potential predictability; this is likely to be higher than what is achieved in practice.

8 Multidecadal behavior

Figure 8 shows the smoothed FO time series, in which subdecadal variability is suppressed. Series for states 1 and 2 do not exhibit marked long-term trends, although decadal variations are evident. Series for states 3 and 4 diverge, however, the former tending upward. This inverse behavior also characterizes decadal variations, and suggests similarities with the interannual case.

Figures 7c and 7d show the first canonical patterns for the smoothed data, which are seen to be similar to those for the interannual series. The first three correlations are also significant (at 0.001) in this case, and explain 52%, 24% and 7% of the station rainfall variance, respectively. The smoothed ISMR is also well-correlated with the first FO canonical variate ($r = 0.85$, p-value of 0.015 for a two-way test on 5 d.o.f.), so an appreciable fraction of the decadal variance can be related to the state frequencies, even though the states themselves are diagnosed with

DOI: 10.1002/qj

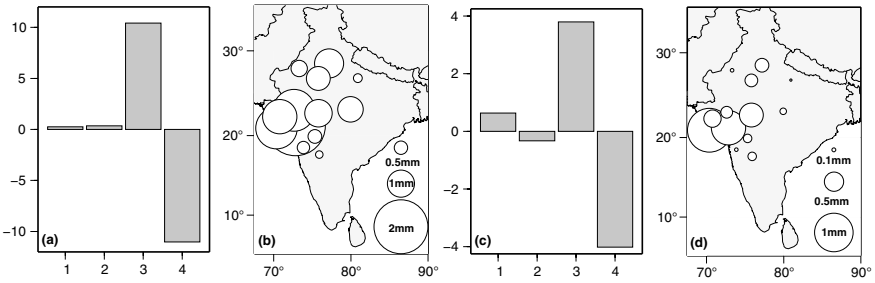

Figure 7. First canonical patterns for (a and c) state FO and (b and d) mean daily rainfall amount. (a) and (b) refer to the subdecadal, (c) and (d) to the lowpassed, series. The single negative value, to the northeast in (d), is shaded.

respect to daily data. Thus, it appears that decadal variations of the ISMR amount in part to an integration of variations in numbers of active and break days. This can be viewed as an extension of the intraseasonal-interannual relation identified by Goswami and Mohan (2001).

9 Discussion

The homogeneous HMM is utilized herein as a diagnostic tool, and provides a compact description of daily rainfall variability over the station network. The relationships detailed, between variations in FO of the diagnosed states, station rainfall distributions and various monsoon features (large-scale atmospheric flow, ISO, all-India monsoon rainfall, ENSO interaction, longer-period variability) indicate that this description contains much information about real physical processes.

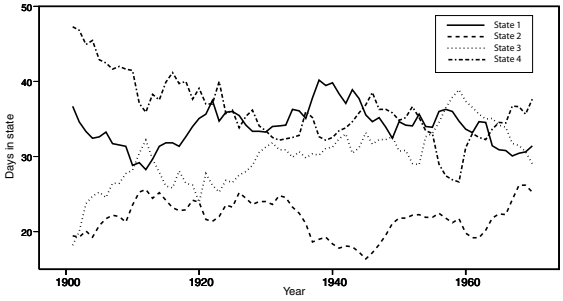

Figure 8. Time evolution of filtered state occurrence frequencies. Series shown are 11-yr moving averages.

Well-defined atmospheric modes corresponding to the states are consistent with both the state rainfall patterns and the large-scale structure of the monsoon. This correspondence may in fact owe something to the fact that the monsoon *is* a large-scale phenomenon, whose modes may be accessible in this way from any similar network meeting minimal requirements for spatial coverage and distribution.

It was shown that year-to-year fluctuations in the first CCA mode, representing inverse variations in the FO of states 3 and 4, play an important role in monsoon variations on even decadal time scales. However, it may also be possible that the HMM has diagnosed rainfall patterns that can be associated with both long- and short-period variability, and thus that would also be recovered from data having lower time resolution to begin with.

10 Summary and Conclusions

A homogeneous hidden Markov model is applied to daily monsoon (JJAS) rainfall data on a network of 13 stations in west-central India for the years 1901-1970. The HMM associates patterns of rainfall received at the stations with a set of hidden states that progress in time as a first-order Markov process. For the diagnostic purposes of the present work, a model having four hidden states was found to be optimal.

DOI: 10.1002/qj

The diagnosed states were found to play distinct roles in the seasonal march of the monsoon, and the associated atmospheric composites to correspond sensibly with state rainfall characteristics. Episodes of dry-state occurrence during the peak rainy season were shown to correspond well with independently diagnosed monsoon breaks, while detailed analysis of the time evolution of "most-likely" states reveals a likely correspondence with phases in the northward propagation of convective disturbances characteristic of the ISO.

On interannual time scales, a strong relationship between ENSO and monsoon rainfall is found for the period under study. Canonical correlation analysis identifies a primary mode in which the occurrence frequencies of the driest and wettest states vary in opposing senses. Both all-India monsoon rainfall and a typical ENSO index are found to project strongly onto this mode, implying that the state frequencies are strongly coupled to both seasonal rainfall totals and ENSO. These relationships persist on decadal time scales, suggesting that long-period shifts in monsoon rainfall can ultimately be linked to interannual variations in state FO. This diagnosis differs from that of Moron et al. (2006) with respect to Senegalese rainfall, in which decadal variability was found to be primarily a consequence of within-state variation, while interannual variability was more strongly influenced by FO.

The potential of the non-homogeneous HMM as a predictive downscaling tool was discussed. A preliminary experiment utilizing the diagnosed FO series as predictors suggested that such a tool may indeed be useful, although better quantification awaits further research. A related application, in an area of research that has received increasing attention of late, is the generation of weather-within-climate data, in the context of long-range climate change studies.

Acknowledgements

We appreciate the helpful advice and comments offered by many staff members at the IRI, including Lisa Goddard, Vincent Moron and Michael Tippet, and by Padhraic Smyth of the University of California, Irvine. This research was supported by Department of Energy grant DE-FG02-02ER63413.

References

Abrol, I., 1996: India's agriculture scenario. *Climate Variability and Agriculture*, Y. Abrol, S. Gadgil, and G. Pant, eds., Narosa, New Delhi, 19–25.

Annamalai, H., J. M. Slingo, K. R. Sperber, and K. Hodges, 1999: The mean evolution and variability of the Asian summer monsoon: Comparison of ECMWF and NCEPNCAR reanalyses. *Mon. Wea. Rev.*, **127**, 1157–1186.

Barnett, T. P. and R. Preisendorfer, 1987: Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Wea. Rev.*, **115**, 1825–1850.

Barnston, A. G., H. M. van den Dool, D. R. Rodenhuis, C. R. Ropelewski, V. E. Kousky, E. A. O'Lenic, R. E. Livezey, S. E. Zebiak, M. A. Cane, T. P. Barnett, N. E. Graham, M. Ji, and A. Leetmaa, 1994: Long-lead seasonal forecasts—where do we stand? *Bull. Amer. Meteor. Soc.*, **75**, 2097–2114.

Bharadwaj, S., 2004: Make economy monsoon-proof: Assocham chief. *Times of India*, http://timesofindia.indiatimes.com/articleshow/796629.cms.

Dempster, A. P., N. M. Laird, and D. R. Rubin, 1977: Maximum likelihood from incomplete data via the *EM* algorithm. *J. R. Stat. Soc.*, **B39**, 1–38.

DOI: 10.1002/qj

Forney, Jr., G. D., 1978: The viterbi algorithm. *Proc. IEEE*, **61**, 268–278.

Gadgil, S., 1995: Climate change and agriculture: An Indian perspective. *Curr. Sci.*, **69**, 649–659.

— 2003: The Indian monsoon and its variability. *Ann. Rev. Earth Planet. Sci*, **31**, 429–467, doi:10.1146/annurev.earth.31.100901.141251.

Gadgil, S. and G. Asha, 1992: Intraseasonal variation of the summer monsoon. Part I: Observational aspects. *J. Meteor. Soc. Japan*, **70**, 517–527.

Gadgil, S. and S. Gadgil, 2002: The Indian monsoon, GDP and agriculture. *Econom. Pol. Weekly*, submitted.

Gadgil, S. and P. V. Joseph, 2003: On breaks of the Indian monsoon. *Proc. Indian Acad. Sci. (Earch Planet. Sci.)*, **112**, 529–558.

Gadgil, S. and K. R. Kumar, 2006: The Asian monsoon — agriculture and economy. *The Asian Monsoon*, B. Wang, ed., Springer, 651–683.

Ghahramani, Z., 2001: An introduction to hidden Markov models and Bayesian networks. *Int. J. Pat. Rec. Art. Int.*, **15**, 9–42.

Ghil, M. and A. Robertson, 2002: "waves" vs. "particles" in the atmosphere's phase space: A pathway to long-range forecasting? *Proc. Nat. Acad. Sci. USA*, **99**, 2493–2500.

Goswami, B. and P. K. Xavier, 2003: Potential predictability and extended range prediction of Indian Summer monsoon breaks. *Geophys. Res. Lett.*, **30**, doi:10.1029/2003GL017810.

Goswami, B. N., 2005: South Asian monsoon. *Intraseasonal Variability in the Atmosphere-Ocean Climate System*, W. K. M. Lau and D. E. Waliser, eds., Springer, Berlin Heidelberg, chapter 2, 19–61.

Goswami, G. N. and R. S. A. Mohan, 2001: Intraseasonal oscillations and interannual variability of the Indian summer monsoon. *J. Climate*, **14**, 1180–1198.

Hastings, D. A. and P. K. Dunbar, 1998: Development & assessment of the global land one-km base elevation digital elevation model (globe). *Int. Soc. Photogramm. Remote Sens. Archives*, **32**, 218–221.

Hughes, J. P. and P. Guttorp, 1994: A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Res. Research*, **30**, 1535–1546.

Hughes, J. P., P. Guttorp, and S. P. Charles, 1999: A non-homogeneous hidden Markov model for precipitation occurrence. *J. Royal Stat. Soc. C*, **48**, 15–30, doi:10.1111/1467-9876.00136.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chellah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471.

Krishnamurthy, V. and B. N. Goswami, 2000: Indian monsoon—ENSO relationship on interdecadal timescale. *J. Climate*, **13**, 579–595, doi:10.1175/1520-0442(2000)013<0579:IMEROI>2.0.CO;2.

Kumar, K. K., B. Rajagopalan, and M. A. Cane, 1999: On the weakening relationship between the

Indian monsoon and ENSO. *Science*, **284**, 2156–2159, doi:10.1126/science.284.5423.2156.

Kumar, K. K., B. Rajagopalan, M. Hoerling, G. Bates, and M. Cane, 2006: Unraveling the mystery of Indian monsoon failure during El Niño. *Science*, **314**, 115–119.

Legates, D. R. and C. J. Willmott, 1990: Mean seasonal and spatial variability in gauge-corrected global precipitation. *Int. J. Climatol.*, **10**, 111–127.

Moron, V., A. W. Robertson, M. N. Ward, and O. N'Diaye, 2006: Weather types and rainfall over Senegal. Part I: Observational analysis. *J. Climate*, submitted.

NCDC, 2002: Data documentation for dataset 9618, Global Summary of the Day. Technical report, National Climatic Data Center (NCDC), National Oceanic and Atmospheric Administration (NOAA).

Rai, S., 2005: Monsoon still helps push India's economy. *Int. Herald Tribune*, http://www.iht.com/articles/2005/06/03/news/india.php.

Rajeevan, M., J. Bhate, J. Kale, and B. Lal, 2006: High resolution daily gridded rainfall data for the Indian region: Analysis of break and active monsoon spells. *Curr. Sci.*, **91**, 296–306.

Rao, Y. P., 1976: Southwest monsoon. India Meteorological Dept., meteorological monograph, 366 pp.

Rasmusson, E. M. and T. H. Carpenter, 1983: The relationship between eastern equatorial Pacific sea surface temperatures and rainfall over India and Sri Lanka. *Mon. Wea. Rev.*, **111**, 517–528.

Robertson, A. W., S. Kirshner, and P. Smyth, 2004: Downscaling of daily rainfall occurrence over northeast Brazil using a hidden Markov model. *J. Climate*, **17**, 4407–4424.

Robertson, A. W., S. Kirshner, P. Smyth, S. P. Charles, and B. C. Bates, 2006: Subseasonal-to-interdecadal variability of the Australian monsoon over North Queensland. *Q.J.R. Meteorol. Soc*, **132**, 519–542.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Shukla, J., 1987: Interannual variability of monsoons. *Monsoons*, J. S. Fein and P. L. Stephens, eds., Wiley, 399–464.

Sontakke, N. A., G. B. Pant, and N. Singh, 1993: Construction of all-India summer monsoon rainfall series for the period 1844-1991. *J. Climate*, **6**, 1897–1811.

Uppala, S. M., 2001: ECMWF ReAnalysis 1957-2001, ERA-40. ERA-40 Project Rep. Series 3, ECMWF, Reading, UK.

Webster, P. J., V. O. Magana, T. N. Palmer, J. Shukla, R. A. Tomas1, M. Yanai, and T. Yasunari, 1998: Monsoons: Processes, predictability, and the prospects for prediction. *J. Geophys. Res.*, **103**, 14,451–14,510.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. Number 91 in International Geophysics Series, Academic Press, 627 pp.

Woolhiser, D. A. and J. Roldán, 1982: Stochastic daily precipitation models 2. a comparison of distiributions of amounts. *Water Resour. Res.*, **18**, 1461–1468.

Xie, P. and P. Arkin, 1996: Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *J. Climate*, **9**, 840–858.