Quarterly Journal of the
Royal Meteorological Society

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

Journal:	*Quarterly Journal of the Royal Meteorological Society*
Manuscript ID:	QJ-07-0008.R2
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Greene, Arthur; International Research Institute for Climate and Society, Columbia University Robertson, Andrew; International Research Institute for Climate and Society, Columbia University Kirshner, Sergey; University of Alberta, Computing Science
Keywords:	Climate diagnosis, Precipitation, Statistical modeling

Q. J. R. Meteorol. Soc. **00**: 1–19 (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[1]* A. W. Robertson[1] and S. Kirshner[2]

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

[2] *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. 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 (e.g., Gadgil, 2003; Webster *et al.*, 1998; Abrol, 1996; Rai, 2005; Gadgil and Kumar, 2006; Gadgil and Gadgil, 2006). In the present study, daily monsoon rainfall at a small network of stations is decomposed using a hidden Markov model (HMM). The HMM is utilized here as a *diagnostic* tool; this is a necessary step if such a model is eventually to be deployed for precipitation downscaling or simulation (Hughes *et al.*, 1999; Bellone *et al.*, 2000). To the best of our knowledge HMMs have not heretofore been deployed in this regional setting; this renders the present investigation both novel and of interest generally, with regard to the diagnostic utility of such models in the monsoon domain.

The HMM associates observed patterns of daily rainfall with a small set of "hidden states," which proceed in

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

time as a first-order Markov process (Hughes and Guttorp, 1994; Norris, 1997). It may be considered a parsimonious description of the raw rainfall observations or, alternatively, as a method of data reduction, by which the essential structural attributes of the complex observational data are represented by a small, therefore more comprehensible, set of parameters.

The HMM also provides a simple means of generating synthetic precipitation series that have some of the statistical properties (including spatial covariance) of the data to which it is fit. Here, however, this is not the goal; in particular, the model employed includes only seasonal-mean transition probabilities, and is thus incapable of simulating the rise and fall of the seasonal cycle. However, once the hidden states are identified, their progression in time can be recovered, and intraseasonal variability thereby *diagnosed*. It is this diagnosis that lies at the core of the present work.

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 without any filtering or gridding, yet is shown to capture characteristic features of rainfall variability across a broad range of time scales. This suggests the existence of some mechanism that links variations occurring on these different scales. It has been noted, for example, that the "active" and "break" phases that characterize subseasonal monsoon variability "add up" to produce interannual variations (Gadgil and Asha, 1992; Gadgil, 1995; Goswami and Xavier, 2003; Goswami, 2005). It is thus of interest to see whether such aggregation might also play a role in decadal monsoon fluctuations.

The link between the monsoon and the 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 have tended to be associated with weak monsoons, at least up until the late 1970s (Kumar *et al.*, 1999). This linkage is explored here through canonical correlation analysis (CCA), applied to the occurrence frequencies of the diagnosed states and the station rainfall data. CCA is then extended, to examine multidecadal variability.

Section 2 describes the datasets employed and Section 3 some climatological characteristics of the data. The HMM is discussed in Section 4, while Section 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 Sections 9 and 10.

2 Data

A 70-year record of daily rainfall 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 1901–70, 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

DOI: 10.1002/qj

Table I. Stations whose records are utilized herein. Columns indicate station number (corresponding to the numbers shown in Fig. 1a), station ID as provided in the dataset, station name, latitude (°N), longitude (°E), climatological rainfall occurrence probability and mean intensity (rainfall amount on wet days, defined as those with amounts ≥ 0.1 mm).

No.	ISTA	Name	Lat	Lon	P(R)	$\bar{I}$
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

in Section 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 Jun–Sep rainfall over approximately 300 stations (Sontakke *et al.*, 1993).

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 climatological rainfall occurrence probabilities and mean daily intensities (rainfall amounts on wet days), respectively, over the network. Topographic contours from the GLOBE digital elevation model (Hastings and Dunbar, 1998) are also shown. Both probabilities and intensities (also given in Table I), are computed conditional on a minimum daily rainfall amount of 0.1 mm, the minimum non-zero value recorded in the dataset.

In Fig. 1, high probability and intensity values along the western coast reflect onshore flow, as shown in Fig. 5a, 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 *patterns* shown on Figs. 1a and 1b are quite consistent with a high-resolution (1°) data set from the India Meteorological Department (IMD, Rajeevan *et al.*, 2006), including the low probabilities and intensities to the north and north-west, (e.g., station 10, lying in arid Rajasthan), the high intensities at stations 6 and 13 in the main monsoon zone, and the south-east to north-west intensity gradient in going from stations 7–9 toward stations 1–4. Absolute amounts do differ somewhat, with the station data showing generally both higher occurrence probabilities and mean daily intensities than enclosing grid boxes in the IMD product. The former may result from the masking of gridded values below 0.1 mm, the latter from gridbox averaging.

DOI: 10.1002/qj

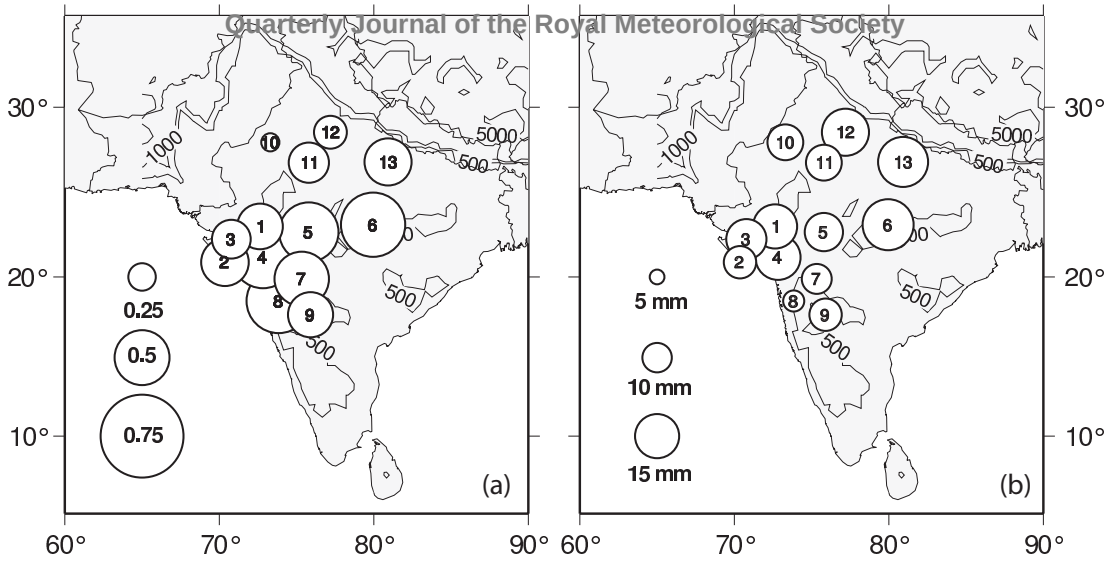

Figure 1. Climatological (a) occurrence probability and (b) mean daily intensity (mm) for Jun–Sep 1901–70.

3.2 Rainfall—seasonal cycle

Figures 2a and 2b illustrate the climatological seasonal cycle for rainfall occurrence frequency 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.

3.3 Circulation

Figure 5a shows mean climatological Jun–Sep horizontal winds at 850 mb and the 500-mb vertical velocity (ω) from the NCEP-NCAR reanalysis, for 1951–70. 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 conditional independence assumptions: 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 (PDF) 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

DOI: 10.1002/qj

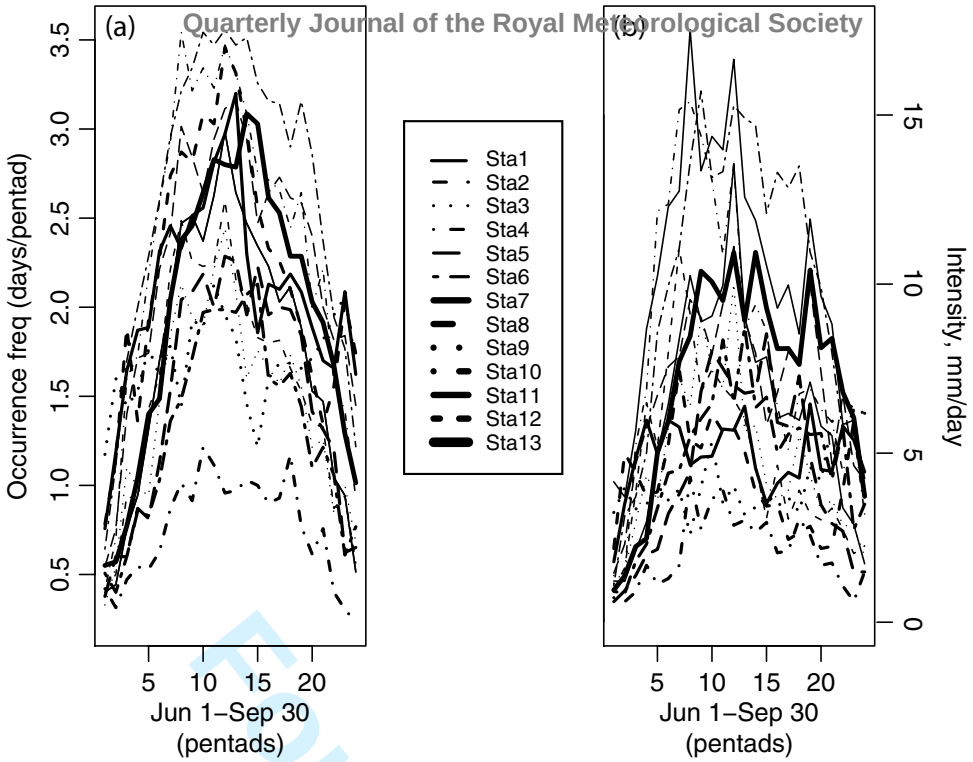

Figure 2. Climatological seasonal cycle of (a) occurrence frequency (days pentad^{-1}) and (b) mean daily intensity (mm) for the 13-station dataset, 1901–70. Occurrence is conditional on a threshold of 0.1 mm.

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

where indices i, m and c refer to state, station and mixture component, respectively, the p_{imc} are weights and t is time. In the summation, $C = 2$, i.e., two exponentials are utilized. Note that while rainfall at each station is characterized by a PDF that is both station- and state-specific, the PDFs for all stations are coupled by state, as per the i subscripts in (1). Thus, the HMM accounts for spatial dependence in the data. There are HMM variants that model spatial dependence in more detail (Kirshner *et al.*, 2004), but we forgo the additional complexity involved in favor of a more easily interpreted model. For a more complete description of the HMM see Robertson *et al.* (2004, 2006).

4.2 Model selection and fitting

The number of states to be modeled must be specified *a priori;* differing objectives may lead to differing choices in this regard. Use of a small number of states facilitates diagnosis and model comprehensibility, the object of this study, while a larger number might be more suitable for the generation of synthetic data. 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 diagnosed states to the propagating convective disturbances characteristic of the intraseasonal oscillation (ISO, see Section 6.3). On the other hand, the five-state model does not add much 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

DOI: 10.1002/qj

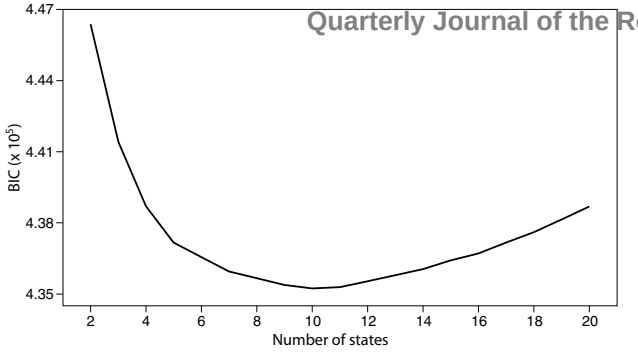

Figure 3. Bayesian information criterion (BIC) for models having differing numbers of hidden states.

Information Criterion (Schwarz, 1978) was applied to fitted models having up to 20 states, (Fig. 3) and indicates that overfitting does not occur when as many as 10 states are 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 was performed by the maximum likelihood approach, using the iterative expectation-maximization (EM) algorithm (Dempster *et al.*, 1977; Ghahramani, 2001). The algorithm was initialized 30 times from random starting points, the run utilized being that with the highest log-likelihood. Estimation was performed using the Multivariate Nonhomogeneous Hidden Markov Model Toolbox, developed by one of the authors (Kirshner, see http://www.cs.ualberta.ca/˜sergey/MVNHMM/).

4.3 Representation in terms of states

Figure 4 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 from the mean values and weights 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 south-west to north-east gradient in mean intensity, while state 2 exhibits a north to south 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, and more importantly for our purposes, 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 transition matrix for the 4-state HMM is shown in Table II. Note that the estimated 3–4 and 4–3 probabilities are both null (actually nonzero, but at least 7 orders of magnitude smaller than the other values in the table); in fact, no direct transitions between these two states occur. This suggests that some *dynamical* process has been encoded by the HMM, such that abrupt changes between the intense endmember conditions represented by these two states are unlikely to occur in nature.

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 from day to day), this feature 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. 6a, the most-likely state sequence (see Section 6.1).

Prepared using qjrms3.cls

DOI: 10.1002/qj

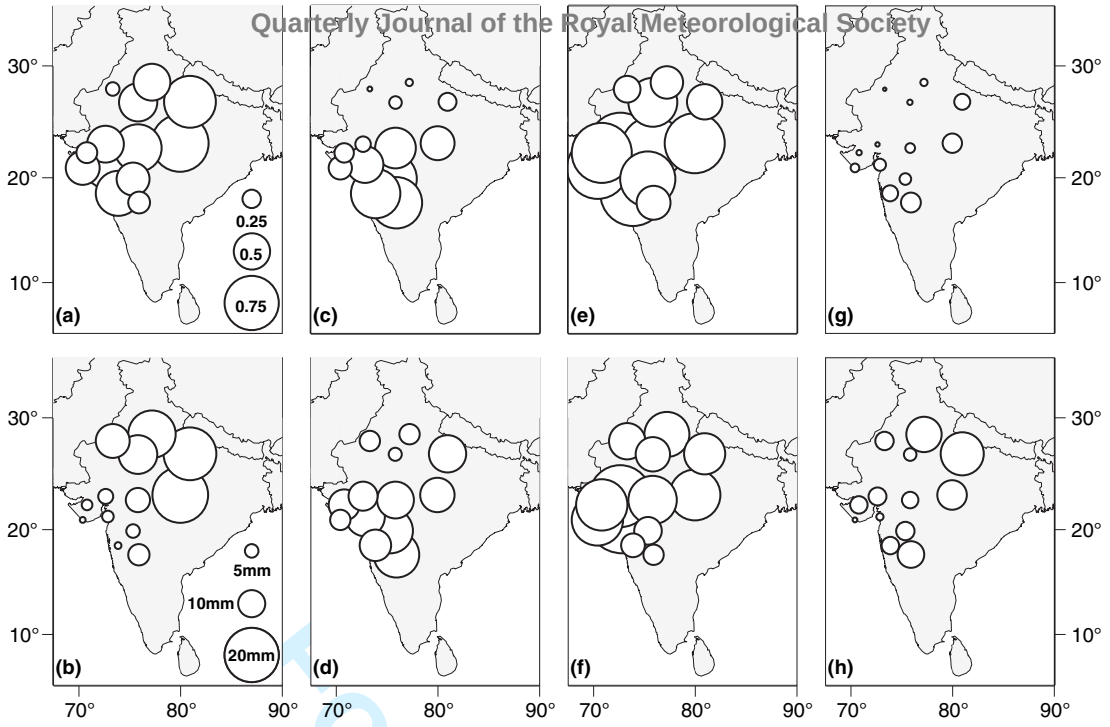

Figure 4. Occurrence probabilities (top) and mean daily intensities (mm) for the 4-state model, states 1–4 shown from left to right. Values are computed conditional on a daily amount of at least 0.1 mm. Legend in leftmost figures applies to all plots in the respective rows.

The values shown in Table II are *time-invariant*, and in essence represent mean transition probabilities for the entire Jun–Sep season. As will be seen, however, there is a pronounced seasonal cycle in the relative frequency of occurrence of the various states, implying that in actuality the probabilities must be temporally modulated. This would clearly present a problem if it was desired to *simulate* the rise and fall of the seasonal cycle, since static transition probabilities can produce only stationary rainfall series. However, once the states have been diagnosed, their relative frequency over the course of the season, and thus the seasonal cycle, as it exists in the *observed* rainfall series, can be determined. This is discussed in more detail in Section 6.

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

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 5b–5e show composited anomalies (with respect to the climatology shown in Fig. 5a) for states 1–4, respectively, for the 4-state HMM.

Generation of these composites requires several steps. First, the raw daily rainfall values must be expressed in terms of the hidden states they represent. This is accomplished here by means of the Viterbi algorithm (see Section 6), which returns a "most-likely" sequence of states, given the state definitions and transition matrix. Each day in the rainfall record is thus identified with its associated state. The days diagnosed as representing each of the states are then collected, with all the days representing state 1 placed in one group, the days representing state

DOI: 10.1002/qj

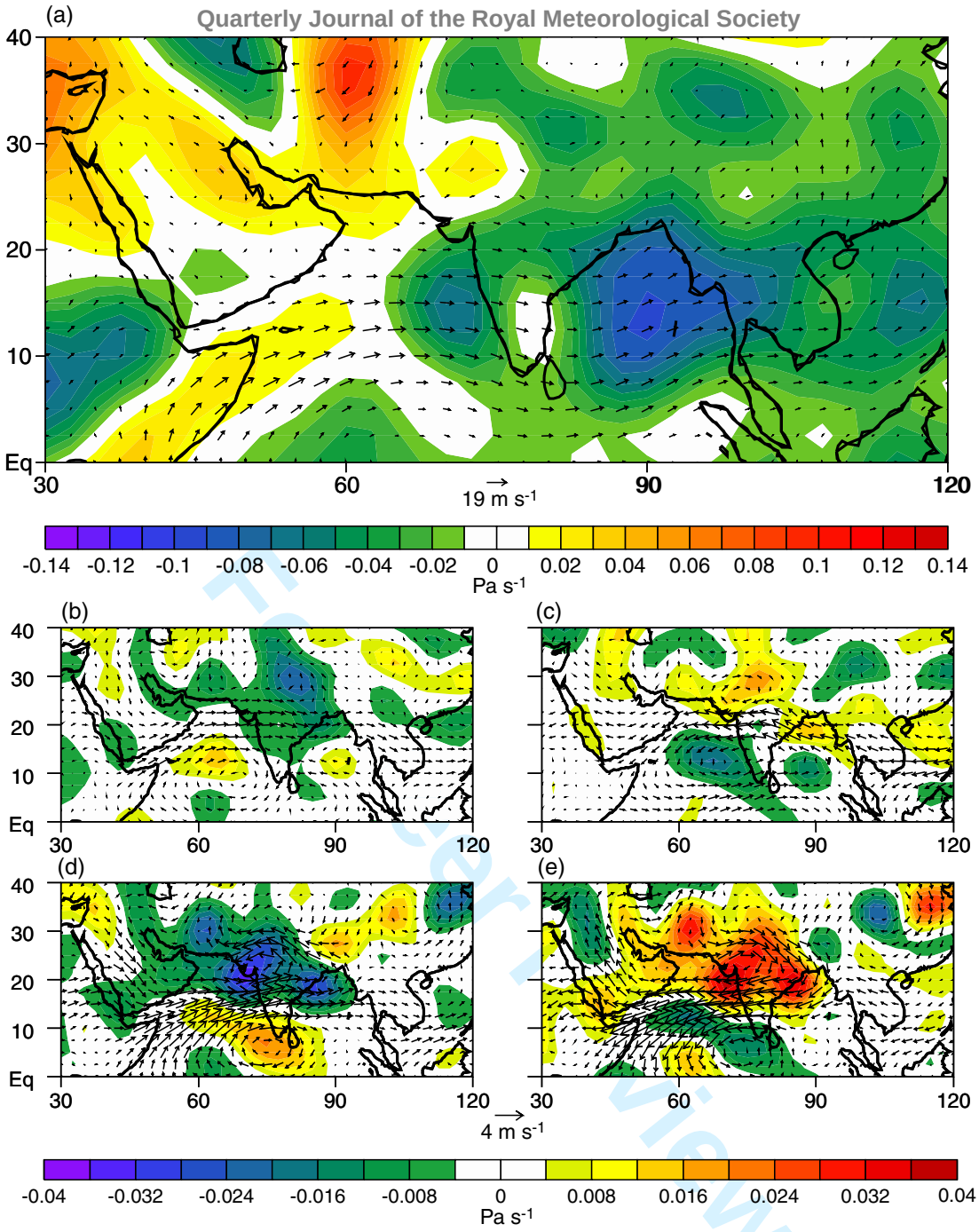

Figure 5. (a) Horizontal winds at 850 mb (vectors, m s^{-1}) and 500-mb vertical velocity (colors, Pa s^{-1}) for the Jun–Sep mean climatology for 1951–70. (b)–(e) Anomalies with respect to (a), for states 1–4, respectively.

2 in a second group, and so on. The composites are formed by taking the arithmetic average of the wind and vertical velocity fields for each of these groups, then subtracting the mean Jul–Sep climatology (shown in Fig. 5a) to produce anomalies.

State 3 (Fig. 5d), 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

DOI: 10.1002/qj

as well as an anticyclonic circulation to the southwest of the Indian peninsula. These features are consistent with the high occurrence probabilities and intensities on the western coast, extending toward the northeast, that characterize this state (Figs. 4e and 4f). 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 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. 5e), the "dry" state, is opposite in sense to 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. 4g and 4h.

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. 5b), 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. 4b), 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. 5c) shows anomalous easterly inflow from the Bay of Bengal, resembling in this sense the dry state. State 2 is also opposed to state 1 (cf. Fig. 5b), 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. 4c and 4d).

Broadly speaking, the circulation anomalies associated with states 1 and 2 can be said to be similar to those associated with 3 and 4, respectively, but weaker, and with the regions of strongest anomalous ascent or descent displaced to the north. States 1 and 3, with anomalous ascending motion over the subcontinent, and in particular the monsoon zone, may both be considered "wet," while state 2, with its anomalous descending motion, shows stronger affinities to the dry state (state 4), most clearly in the northern part of the domain.

Among the questions that naturally arise from inspection of these figures is that of the relationship between the diagnosed states and the wet and dry intraseasonal-scale phenomena known as "active" and "break" phases of the monsoon. This question is taken up in Section 6, along with some other aspects of intraseasonal variability, as viewed through the prism of state decomposition. The rainfall patterns associated with the diagnosed states appear, in any event, to correspond quite sensibly with known large-scale monsoon-related circulation regimes (Gadgil, 2003).

DOI: 10.1002/qj

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. 6a. Figure 6b shows the climatological sequence for 1901–70, accumulating days-in-state over the 70 years.

Figures 6a and 6b reveal a systematic progression in state occurrence over the course of the monsoon season. During the first half of June, state 4, the dry state, dominates, while during the core of the rainy season states 1 and 3 assume primary importance. State 2 plays a quasi-transitional role, first appearing as a bridge between dry and wet conditions in late June, almost disappearing during the wettest part of the season, then returning in September, with increasing representation toward the end of that month. After mid-September the dry state once again becomes dominant. Figure 6b also reveals a subtle evolution of precipitation patterns during the core rainy season, with July favoring state 3 but a shift 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 Jun–Sep season, with standard deviations 10.3, 7.9, 11.8 and 15.2 days, indicating considerable interannual variability. Variability on longer time scales is also suggested by Fig. 6a.

6.2 Monsoon breaks

Figures 6a and 6b both show clearly the dominance of state 4 during the early and late stages of the monsoon. This state also occurs sporadically during the Jul–Aug core of the rainy season, however, suggesting a possible association with monsoon breaks. Gadgil and Joseph (2003) provide a listing of breaks for 1901–89, 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 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 Jul–Aug for 1901–70, 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. 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. A bootstrap test indicated that this result has not been biased on account of serial autocorrelation; we therefore conclude that there exists a high degree of correspondence between occurrence of the HMM dry state and monsoon break days.

DOI: 10.1002/qj

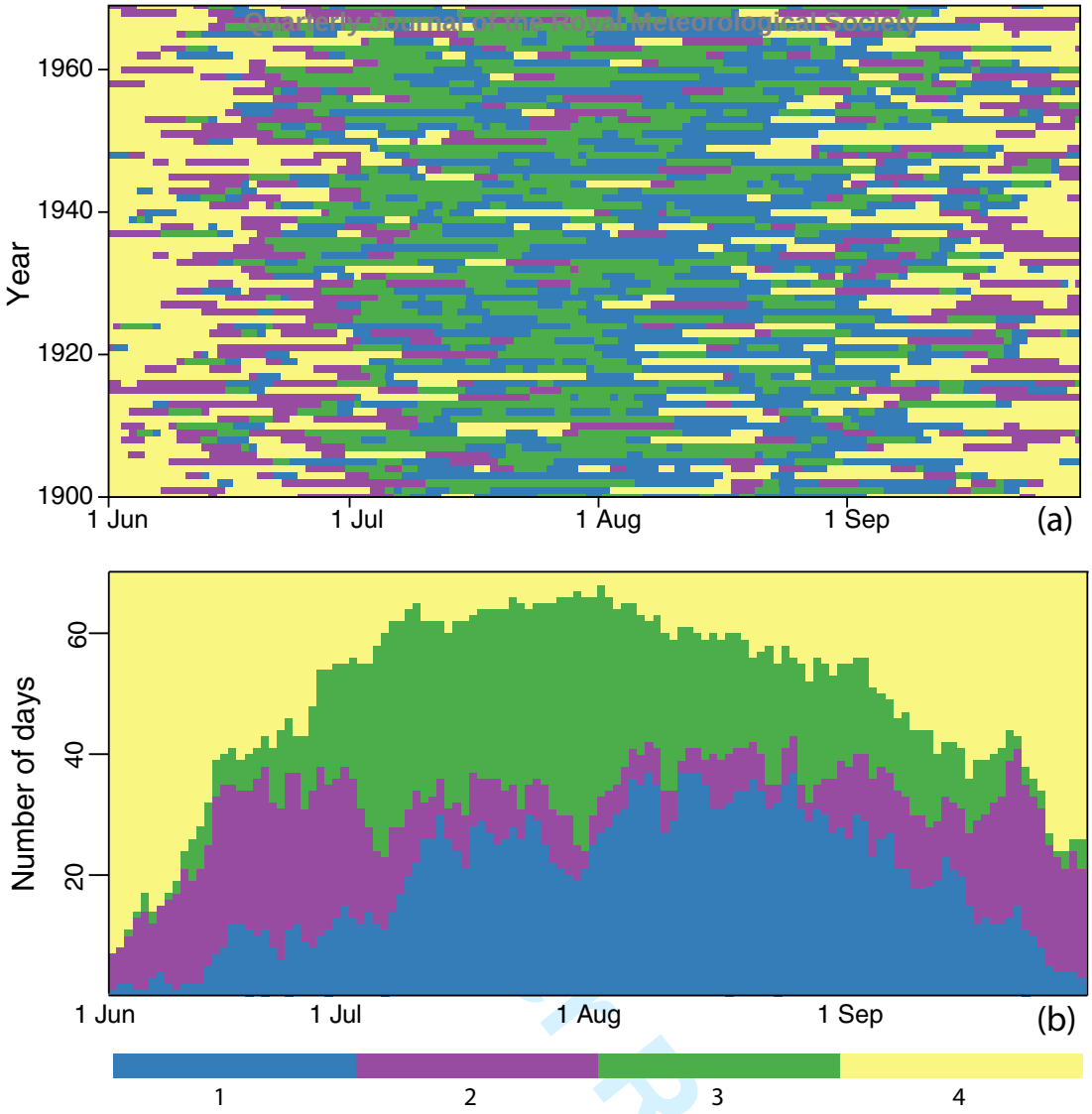

Figure 6. (a) Viterbi sequence of most likely states, 1901–70. (b) Corresponding daily climatology, accumulating days-in-state over the 70-year data period. Colors indicate states 1–4.

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. 5b). Rainfall occurrence probabilities for state 1, however, are uniformly moderate to high (Fig. 4a), quite different from those of the dry state.

Clues to this conundrum may be found in the amount distribution, shown in Fig. 4b, and, somewhat more cryptically, in the Viterbi sequence (Fig. 6a). In the former, 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 preceded by state 1, despite the comparable prevalence of state 3 during Jul–Aug (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

DOI: 10.1002/qj

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 Section 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.

In light of this description, states 1 and 3 (Figs. 5b and 5d) 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 relative frequency with 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, are thus more likely to be captured by the state descriptions.

6.4 Propagation of convective disturbances

We focus here on the Jul–Aug core of the wet season. During these months the monsoon is fully active, the dry periods at the beginning of June and end of September being excluded. Transition probabilities for Jul–Aug, estimated from the Viterbi sequence, are shown in Table III (cf. Table II, which applies to the entire Jun–Sep season, and where transitions to the dry state from states 1 and 2 are considerably more likely). We consider the *off-diagonal* elements in this array, from which most-likely sequences of states may be deduced. Exclusion of elements on the main diagonal is equivalent to considering only transitions from one state to a *different* state, thus ignoring self-transitions. Attention is thereby directed to

Prepared using qjrms3.cls

DOI: 10.1002/qj

Table III. Jul–Aug transition probabilities.

		"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

the temporal patterns of *intraseasonal* variability, rather than the daily transitions.

The most-likely sequence, thus defined, varies according to which state is taken as the starting point, but if we think of the ISO as described by Goswami and Mohan, i.e., as an alternation between two centers of convective activity (with propagation from south to north), we can think of a complete "cycle" as extending from break to break—a break occurring when the locus of convection lies to the south of the equator. Beginning with a break (state 4), the most-likely state sequence is then 4–2–3–1. Figure 7 shows composites of 850-mb relative vorticity corresponding to the wind fields of Figs. 5b-5e. Viewed in the 4–2–3–1 sequence, the plots show a northward progression of the band of positive vorticity, beginning, in state 4, at the southern extremity of the subcontinent. This would be consistent with the northward-propagating disturbances described by Goswami and Mohan (2001).

The Markov chain, of course, follows some *mixture* of all the paths permitted by the transition matrix; thus, there is considerable stochastic variability in the actual progression of states. Nevertheless, the 4–2–3–1 pattern is frequently found intact in the Viterbi sequence.

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 have not been filtered to retain only ISO-band variability, and thus contain information about all time scales.

6.5 Other aspects of intraseasonal variability

From Table III it can be seen that another "preferred" sequence consists of an alternation between states 1 and 3, and also 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 Jul–Aug, and the less-frequent excursions to state 4 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).

A feature of interest in Fig. 6b involves the shift in dominance, during the peak Jul–Aug 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.

7 Interannual variations—Influence of ENSO

The four-state model comprises two "wet" and two "dry" states, with states 3 and 4 the more intense in these two categories, respectively, and 1 and 2 the more attenuated. Over the course of a full season, the number of days spent in each of the states can thus signal relatively wet or dry years; the unfolding in time of these variations constitutes what we would call interannual variability, but now expressed in terms of frequency of occurrence (FO) of the model's hidden states. These occurrence frequencies, which apply to the station network as whole, may in turn be thought of as representing interannual variations in

DOI: 10.1002/qj

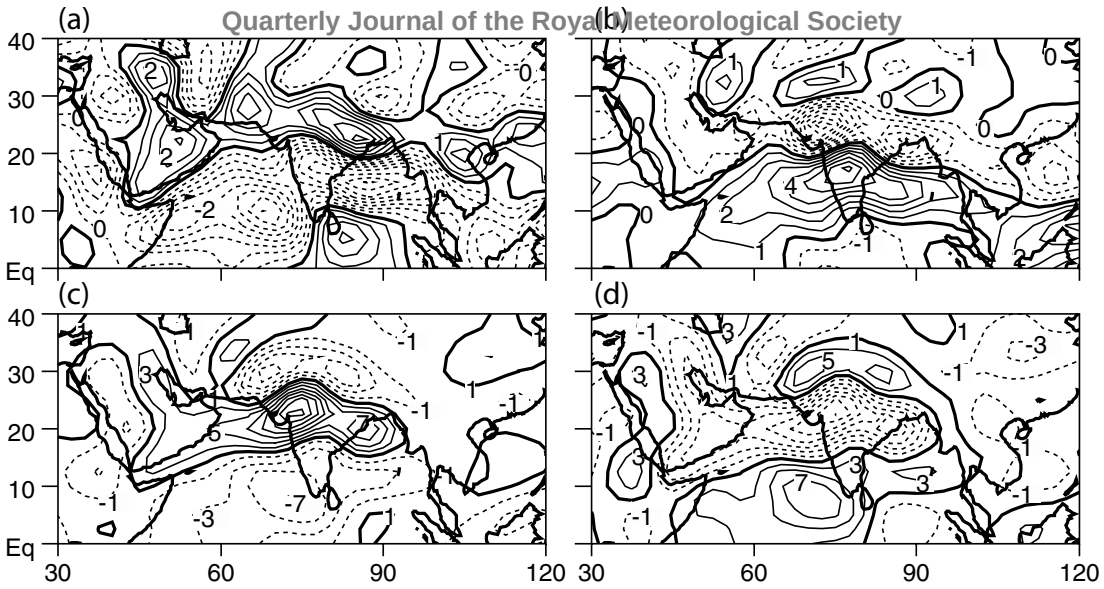

Figure 7. Relative vorticity anomaly composites for the 4-state model, states 1–4 shown in panels (a)–(d), respectively. Units are 10^{-6} s^{-1}.

the large-scale circulation. (Indeed, this has been demonstrated in Section 5.) On the other hand, the states are related to the station rainfall through the structure of the HMM. Thus, state FO links the large spatial scale of the circulation fields with the small scales of station rainfall. This linkage is explored in what follows.

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 (Barnston *et al.*, 1994) 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 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. These relationships confirm the large-scale character of state FO, as would be expected from the results of Section 5.

The relationship between FO and station rainfall cannot be considered for each of the states separately, because FO need not (indeed, cannot) vary independently among states. In addition, there exists the possibility that within-state variation (changes in the character of the states), if systematic, could cause station rainfall variations to diverge from what variations in FO alone would lead us to expect. Canonical correlation analysis (CCA, see, e.g., Wilks, 2006) offers a means of addressing these potentially confounding aspects of the FO-rainfall linkage, and is thus employed here in order to characterize that 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.* (2007) have performed a similar analysis, as part of an investigation of Senegalese rainfall.

DOI: 10.1002/qj

ues, are the state FO series and the mean daily station 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 means. These smoothed versions are then subtracted from the original series, leaving the shorter-period variations as a residual. The Kolmogorov-Smirnov test did not lead to a rejection of the null hypothesis of normality for any of the resulting state or station series; CCA was thus applied without any transformation of variables.

Figures 8a and 8b 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 CCA 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. From the HMM perspective, then, ENSO modulates monsoon rainfall through the agency of the state frequencies, producing lower (higher) counts for state 3 (4) in El Niño years, 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.

Potential 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, which will not be the case in practice.

8 Multidecadal behavior

Figure 9 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 trend in opposite directions, however, the former increasing. This tendency, of states 3 and 4 to vary in opposite senses, also characterizes decadal variations, and suggests similarities with the interannual case.

Figures 8c and 8d 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

DOI: 10.1002/qj

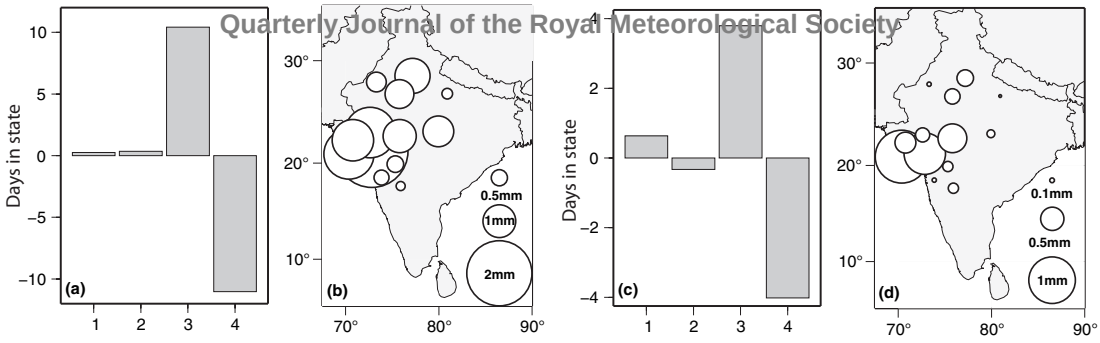

Figure 8. First canonical patterns for state FO (a and c, units are days in state) and mean daily rainfall amount (b and d, units are mm). (a) and (b) refer to the subdecadal, (c) and (d) to the lowpassed, series. The single negative value (most easterly station) in (d), is shaded.

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 respect to daily data. Thus, it appears that decadal variations of the ISMR amount in part to an aggregation, over many years, of wet and dry states. 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.

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 owe something to the fact that the monsoon *is* a large-scale phenomenon, whose modes might be accessible in this way from any similar network meeting some minimal sampling requirement.

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. The possibility that there are, in addition, low-frequency modes of variability whose expression is similar to the behavior that is here attributed to the aggregation of wet and dry states over decade-length periods cannot be ruled out. However, such modes are not amenable to discovery through the agency of the HMM.

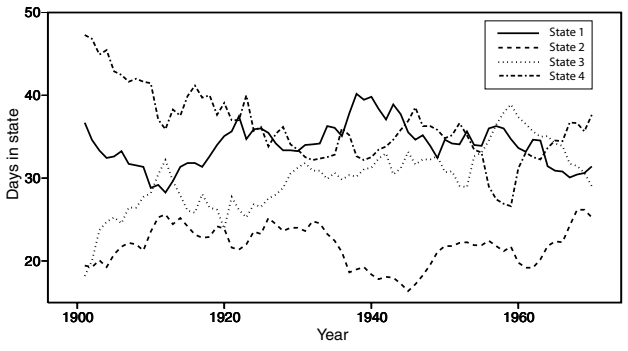

Figure 9. Time evolution of filtered state occurrence frequencies. Series shown are 11-yr moving averages. Units are days-in-state during the 122-day Jul–Sep monsoon season.

10 Summary and Conclusions

A homogeneous hidden Markov model is applied to daily Indian monsoon rainfall on a network of 13 stations in

DOI: 10.1002/qj

central India, for the years 1901–70. The HMM variations in state FO. This diagnosis, which has not been 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 purposes of the present work, a model having four hidden states is found to be optimal, in that it captures sufficient detail to represent essential features of monsoon variability, while retaining adequate interpretive simplicity for the purposes of the present exposition. To the best of our knowledge, application of a statistical model of this type in the Indian monsoon domain has not previously been attempted.

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. This evidence lends credence to the HMM representation of monsoon spatiotemporal variability, and suggests that such models may also find use in other monsoon-dominated circulation regimes.

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

made previously, differs from that of Moron *et al.* (2007) 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.

A preliminary experiment utilizing the diagnosed FO series as predictors suggested that the HMM may prove useful in this regional setting as a statistical downscaling tool, although better quantification awaits further investigation. 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. The model validation presented here represents an important step toward the realization of these applications.

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 US Department of Energy grant DE-FG02-02ER63413.

References

Abrol I. 1996. India's agriculture scenario. In: *Climate Variability and Agriculture*, Abrol Y, Gadgil S, Pant G (eds), 424 pp., Narosa: New Delhi, pp. 19–25.

Annamalai H, Slingo JM, Sperber KR, Hodges K. 1999. The mean evolution and variability of the Asian summer monsoon: Comparison of ECMWF and NCEP-NCAR reanalyses. *Mon. Weather Rev.* **127**: 1157–1186.

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

DOI: 10.1002/qj

Barnston AG, van den Dool HM, Rodenhuis DR, Ropelewski CF, Kousky VE, O'Lenic EA, Livezey RE, Zebiak SE, Cane MA, Barnett TP, Graham NE, Ji M, Leetmaa A. 1994. Long-lead seasonal forecasts—where do we stand? *Bull. Am. Meteorol. Soc.* **75**(11): 2097–2114.

Bellone E, Hughes JP, Guttorp P. 2000. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Clim. Res.* **15**: 1–12.

Dempster AP, Laird NM, Rubin DR. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *J. R. Stat. Soc.* **B39**: 1–38.

Forney, Jr GD. 1978. The Viterbi algorithm. *Proc. IEEE* **61**: 268–278.

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

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

Gadgil S, Gadgil S. 2006. The Indian monsoon, GDP and agriculture. *Econom. Pol. Weekly* **41**(47): 4887–4895.

Gadgil S, Joseph PV. 2003. On breaks of the Indian monsoon. *Proc. Indian Acad. Sci. (Earth Planet. Sci.)* **112**(4): 529–558.

Gadgil S, Kumar KR. 2006. The Asian monsoon — agriculture and economy. In: *The Asian Monsoon*, Wang B (ed), 844 pp., Springer: Berlin, pp. 651–683.

Ghahramani Z. 2001. An introduction to hidden Markov models and Bayesian networks. *Int. J. Pattern Recogn. Artif. Intell.* **15**(1): 9–42, doi:10.1142/S0218001401000836.

Ghil M, Robertson A. 2002. "Waves" vs. "particles" in the atmosphere's phase space: A pathway to long-range forecasting? *Proc. Nat. Acad. Sci* **99**(suppl. 1): 2493–2500.

Goswami B, Xavier PK. 2003. Potential predictability and extended range prediction of Indian summer monsoon breaks. *Geophys. Res. Lett.* **30**(18), doi:10.1029/2003GL017810.

Goswami BN. 2005. South Asian monsoon. In: *Intraseasonal Variability in the Atmosphere-Ocean Climate System*, Lau WKM, Waliser DE (eds), ch. 2, 436 pp., Springer: Berlin, pp. 19–61.

Goswami GN, Mohan RSA. 2001. Intraseasonal oscillations and interannual variability of the Indian summer monsoon. *J. Climate* **14**: 1180–1198.

Hastings DA, Dunbar PK. 1998. Development and assessment of the global land one-km base elevation digital elevation model (GLOBE). *Int. Soc. Photogramm. Remote Sens. Arch.* **32**(4): 218–221.

Hughes JP, Guttorp P. 1994. A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resour. Res.* **30**(5): 1535–1546.

Hughes JP, Guttorp P, Charles SP. 1999. A non-homogeneous hidden Markov model for precipitation occurrence. *J. Royal Stat. Soc. C* **48**(1): 15–30, doi:10.1111/1467-9876.00136.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chellah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D. 1996. The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**(3): 437–471.

Kirshner S, Smyth P, Robertson AW. 2004. Conditional Chow-Liu tree structures for modeling discrete-valued vector time series. Technical Report 04-04, School of Information and Computer Science, University of California, Irvine.

Krishnamurthy V, Goswami BN. 2000. Indian monsoon-ENSO relationship on interdecadal timescale. *J. Climate* **13**(3): 579–595, doi: 10.1175/1520-0442(2000)013<0579:IMEROI>2.0.CO;2.

Kumar KK, Rajagopalan B, Cane MA. 1999. On the weakening relationship between the Indian monsoon and ENSO. *Science* **284**(5423): 2156–2159, doi:10.1126/science.284.5423.2156.

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

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

Moron V, Robertson AW, Ward MN, Ndiaye O. 2007. Weather types and rainfall over Senegal. part I: Observational analysis. *J. Climate* In press.

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). Available online at http://www4.ncdc.noaa.gov/ol/documentlibrary/datasets.html.

Norris J. 1997. *Markov chains*. Cambridge University Press: Cambridge. 237 pp.

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, Bhate J, Kale J, Lal B. 2006. High resolution daily gridded rainfall data for the Indian region: Analysis of break and active monsoon spells. *Current Sci.* **91**(3): 296–306.

Rao YP. 1976. Southwest monsoon. India Meteorological Dept. Meteorological Monograph, 366 pp.

Rasmusson EM, Carpenter TH. 1983. The relationship between eastern equatorial Pacific sea surface temperatures and rainfall over India and Sri Lanka. *Mon. Weather Rev.* **111**: 517–528.

Robertson AW, Kirshner S, Smyth P. 2004. Downscaling of daily rainfall occurrence over northeast Brazil using a hidden Markov model. *J.*

http://mc.manuscriptcentral.com/qj
DOI: 10.1002/qj

mate **17**: 4407–4424.

Robertson AW, Kirshner S, Smyth P, Charles SP, Bates BC. 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**(6): 461–464.

Shukla J. 1987. Interannual variability of monsoons. In: *Monsoons*, Fein JS, Stephens PL (eds), 632 pp., Wiley: New York, pp. 399–464.

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

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

Webster PJ, Magana VO, Palmer TN, Shukla J, Tomas1 RA, Yanai M, Yasunari T. 1998. Monsoons: Processes, predictability, and the prospects for prediction. *J. Geophys. Res.* **103**(C7): 14,451–14,510.

Wilks DS. 2006. *Statistical methods in the atmospheric sciences*. No. 91 in: the International Geophysics Series, Academic Press: Burlington, MA, second edn.

Woolhiser DA, Roldán J. 1982. Stochastic daily precipitation models 2. A comparison of distributions of amounts. *Water Resour. Res.* **18**(5): 1461–1468.

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

DOI: 10.1002/qj