Hidden Markov models for modeling daily rainfall occurrence over Brazil

Technical Report UCI-ICS 03-27
Information and Computer Science
University of California, Irvine

Andrew W. Robertson
International Research Institute for Climate Prediction (IRI)
The Earth Institute at Columbia University
Monell 230, 61 Route 9W,
Palisades, NY 10964
awr@iri.columbia.edu

Sergey Kirshner and Padhraic Smyth
School of Information and Computer Science
University of California, Irvine
Irvine, CA 92697-3425
{skirshne,smyth}@ics.uci.edu

November 24, 2003

Abstract

A hidden Markov model (HMM) is used to describe daily rainfall occurrence at ten gauge stations in the state of Ceará in northeast Brazil during the February–April wet season 1975–2002. The model assumes that rainfall occurrence is governed by a few discrete states, with Markovian daily transitions between them. Four "hidden" rainfall states are identified. One pair of the states represents wet vs. dry conditions at all stations, while a second pair of states represents north-south gradients in rainfall occurrence. The estimated daily state-sequence is characterized by a systematic seasonal evolution, together with considerable variability on intraseasonal, interannual and longer time scales. The first pair of states are shown to be associated with large-scale displacements of the tropical convergence zones, and with teleconnections typical of the El Niño-Southern Oscillation and the North Atlantic Oscillation. A trend toward greater rainfall occurrence in the north of Ceará compared to the south since the 1980s is identified with the second pair of states.

A non-homogeneous HMM (NHMM) is then used to downscale daily precipitation occurrence at the ten stations, using general circulation model (GCM) simulations of seasonal-mean large-scale precipitation, obtained with historical sea surface temperatures prescribed globally. Interannual variability of the GCM's large-scale precipitation simulation is well correlated with seasonal- and spatial-averaged station rainfall-occurrence data. Simulations from the NHMM are found to be able to reproduce this relationship. The GCM-NHMM simulations are also able to capture quite well interannual changes in daily rainfall occurrence and 10-day dry spell frequencies at some individual stations. It is suggested that the NHMM provides a useful tool (a) to understand the statistics of daily rainfall occurrence at the station level in terms of large-scale atmospheric patterns, and (b) to produce station-scale daily rainfall sequence scenarios for input into crop models etc.

1 Introduction

One of the major challenges in tailoring seasonal climate forecasts to meet societal needs, is that the potential users of climate information are often concerned with the characteristics of high-frequency weather at a particular location. Unfortunately, the statistics of local weather are generally poorly represented in the coarse-resolution general circulation models (GCMs) that are typically used to make seasonal climate forecasts. Moreover, the seasonal predictability of high-frequency local information is often in serious doubt. As part of the endeavor to produce useful seasonal climate forecasts, an important task is to understand, on a regional basis, just what aspects of the "weather-within-climate" are predictable a season or more in advance. While the deterministic predictability limit of weather is on the order of 10 days, some boundary-forced longer-term predictability of weather *statistics* may often exist—termed predictability of the "second kind" by Lorenz (1963).

Northeast (NE) Brazil is a region with a high potential for seasonal rainfall predictability during the February–April rainy season, due to strong teleconnections with the El Niño-Southern Oscillation (ENSO) and with variability of the tropical Atlantic Ocean (Hastenrath and Heller 1977, Moura and Shukla 1981, Nobre and Shukla 1996). At the seasonal-mean scale, some of this forecast potential has been realized in the seasonal forecasts made at the International Research Institute for Climate Prediction (IRI) during the 1998–2001 period (Goddard et al. 2003). Some of the highest rainfall skills on the globe occurred over NE Brazil during this short time interval since the advent of routine IRI seasonal forecasts. The state of Ceará in NE Brazil is also the target of a water-resource management project[1] whose goal is to incorporate climate forecast information into sectorial decision making, for which the "downscaling" potential from GCM simulations to local daily rainfall is an important issue.

The first goal of this paper is to examine the probability distribution of local daily rainfall-occurrence in the state of Ceará on timescales of intraseasonal to interannual, and to identify relationships with large-scale atmospheric circulation patterns. Our approach is based on fitting a Hidden Markov Model (HMM) to rainfall records on a network of 10 rain-gauge stations for the period 1975–2002. The HMM is a doubly stochastic model in which the daily probability of local rainfall occurrence (typically on a spatial network of stations) is conditioned on a small number of discrete ("hidden") weather states, with Markovian daily transitions between them. It is a data-driven model, in which the parameters defining each state and the estimated daily sequence of states are derived from a historical record of daily rainfall occurrence. The concept of discrete weather states has a long history in synoptic (Bauer 1951) as well as theoretical midlatitude meteorology (Charney and DeVore 1979). No such literature exists for the tropics, however. We seek to establish a physical basis for the predictability of local daily rainfall statistics over NE Brazil by constructing relationships between the HMM's hidden rainfall states and the large-scale modes of atmosphere-ocean variability associated with ENSO and tropical Atlantic variability (TAV).

For the downscaling of rainfall, Hughes and Guttorp (1994) pioneered the use of the *non-homogeneous* HMM (NHMM) in which the transition probabilities between states are not held fixed, as they are in the classical "homogeneous" HMM, but are allowed to evolve in time according to a small number of large-scale atmospheric exogenous predictor variables. The NHMM links the local rainfall at a network of stations to large-scale atmospheric variables, using the hidden

[1] Information can be found at `http://iri.columbia.edu/application/region/america/Ceara/`

weather states as intermediaries. In the context of downscaling of seasonal forecasts, it may be most meaningful to allow the daily transition probabilities between states to be a function of the GCM's seasonal-average predictions. This would be tantamount to assuming that the information content of the GCM is limited to the seasonal average; it provides a benchmark against which more complex hypotheses can be tested involving atmospheric forcing variables derived from the GCM's daily output.

Northeast Brazil is located near the boundary between two large-scale tropical convergence zones—the Atlantic inter-tropical convergence zone (ITCZ) to the north, and the South American monsoon system (SAMS) centered to the west—and the influence of the South Atlantic subtropical anticyclone situated to the southeast. The rainfall over NE Brazil is highly seasonal and is sensitive to anomalies in the extent of any one of these three phenomena (Hastenrath and Heller 1977). It is semi-arid for much of the year, falling under the influence of the western fringe of the South Atlantic subtropical anticyclone. The principal rainfall season occurs in February-March-April (FMA), when the Atlantic ITCZ reaches its southernmost position and directly overlies NE Brazil, merging with the SAMS near the mouth of the Amazon, as the SAMS (which peaks in austral summer) retreats northward. The time lag in the seasonal migration of the maritime ITCZ is caused by the large thermal heat capacity of the underlying ocean, together with the impact of the continental convection associated with the SAMS which tends to weaken the ITCZ during FMA.

Interannual variability of rainfall is large, and takes the form of droughts that occur when the usual southward seasonal migration of the ITCZ fails to occur, and large rainfall when the latter is amplified. The interannual behavior of the Atlantic ITCZ during boreal spring is closely tied to two inter-related factors: (a) anomalies in the Walker Circulation associated with ENSO events over the tropical Pacific, and (b) changes in the meridional sea surface temperature (SST) gradient in the equatorial Atlantic—the so-called Atlantic "meridional mode"—that may or may not be associated with ENSO (Chiang et al. 2002). The largest ITCZ displacements occur in ENSO years in which pre-existing Atlantic SST anomalies are such as to amplify the direct impact of ENSO (Giannini et al. 2003). The North Atlantic Oscillation (NAO)—which peaks in boreal winter—is one factor that influences the tropical Atlantic meridional mode. The NAO is an intrinsic mode of the atmosphere with a time scale of about a week (Feldstein 2000), with little seasonal predictability. GCM modeling studies suggest that the influence of North Atlantic SST anomalies is small (Hurrell et al. 2002), although some seasonal predictability may arise from stratospheric coupling (Thompson et al. 2002). The meridional mode is influenced from the south as well, and Rossby wave energy on sub-monthly time scales emanating from the South Pacific storm track can reach NE Brazil even during the austral summer (Liebmann et al. 1999).

GCM forecasts of the seasonal average precipitation and temperature are currently made on a routine basis at IRI and at other centers, issued as three-month seasonal averages on the grid-scale of the atmospheric GCMs (typically about 300 km). Experimental downscaled seasonal forecasts are now also being made on an ongoing basis for NE Brazil using high-resolution regional dynamical models (60-km grid) (Sun et al. 2003). Preliminary results are encouraging. However, these high-resolution dynamical models also suffer from model biases and are computationally expensive, motivating the use of statistical approaches.

Having established that certain HMM rainfall states are strongly tied to the large-scale atmospheric circulation, the paper's second goal is to determine to what extent the nonhomogeneous HMM (NHMM) can be used to downscale GCM seasonal climate forecasts over Ceará. The remainder of the paper is laid out as follows. Section 2 describes the observed daily rainfall dataset and the GCM used in the downscaling experiment. The homogeneous HMM is described briefly in Section 3. The hidden states of rainfall occurrence derived from the HMM are presented in Section 4 and discussed in terms of concurrent atmospheric conditions. The results of the GCM downscaling experiment using a non-homogeneous HMM are described in Section 5, with details of the parameter fitting procedure for the NHMM given in Appendix A. A summary and conclusions are presented in Section 6.

2 Observed Rainfall Data and GCM

We use daily rainfall data collected at 10 stations from the state of Ceará in NE Brazil over the years 1975–2002, provided by FUNCEME (Fig. 1). These 10 stations have the longest and most complete reliable records. A wet day is defined as having measurable rainfall. The mean seasonal cycle of rainfall occurrence is plotted in Fig. 2. We select the 90-day period beginning February 1 (February-March-April, FMA), corresponding to the peak rainy season over NE Brazil, and retain the seasonal cycle of rainfall occurrence within the FMA season. The years 1976, 1978, 1984 and 1986 were omitted because of missing data for certain months at one or more stations, yielding 24 complete 90-day seasons (2160 days).

The climatological values of rainfall occurrence probability are plotted at each station on a map of topography in Fig. 1. The largest values occur in the northwest decreasing southward, consistent with the large-scale rainfall pattern associated with the ITCZ and the SAMS. Local topography also appears to play a role in the large probabilities at stations 9 and 10, and perhaps 3 and 6.

The GCM is the ECHAM 4.5 model (Roeckner and Coauthors 1996), for which an ensemble of 24 integrations was available, with historical SSTs prescribed at the lower boundary from the NCEP/NCAR reanalysis data set (Kalnay et al. 1996). Each ensemble member differs only in its initial condition.

3 The Homogeneous Hidden Markov Model (HMM)

Adopting similar notation to that in Hughes et al. (1999), let $\mathbf{R}_t = \left(R_t^1, \ldots, R_t^M\right)$ be a multivariate random vector of rainfall occurrences for a network of M rain stations. Let the observed value $r_t^i = 1$ if rain is observed on day t at station i and $r_t^i = 0$ if it is dry. Let S_t be the hidden rainfall state for day t, taking on one of K values from 1 to K. By $\mathbf{R}_{1:T}$ and $S_{1:T}$ we will denote daily sequences of precipitation occurrences and hidden rainfall states.

An HMM for rainfall data makes two conditional independence assumptions (e.g., see Hughes and Guttorp 1994). The first assumption is that the multivariate precipitation observations $\mathbf{R}_t$ at time t are independent of all other variables in the model up to time t, conditional on the weather state S_t at time t , i.e.,

$$P\left(\mathbf{R}_t | S_{1:t}, \mathbf{R}_{1:t-1}\right) = P\left(\mathbf{R}_t | S_t\right). \tag{1}$$

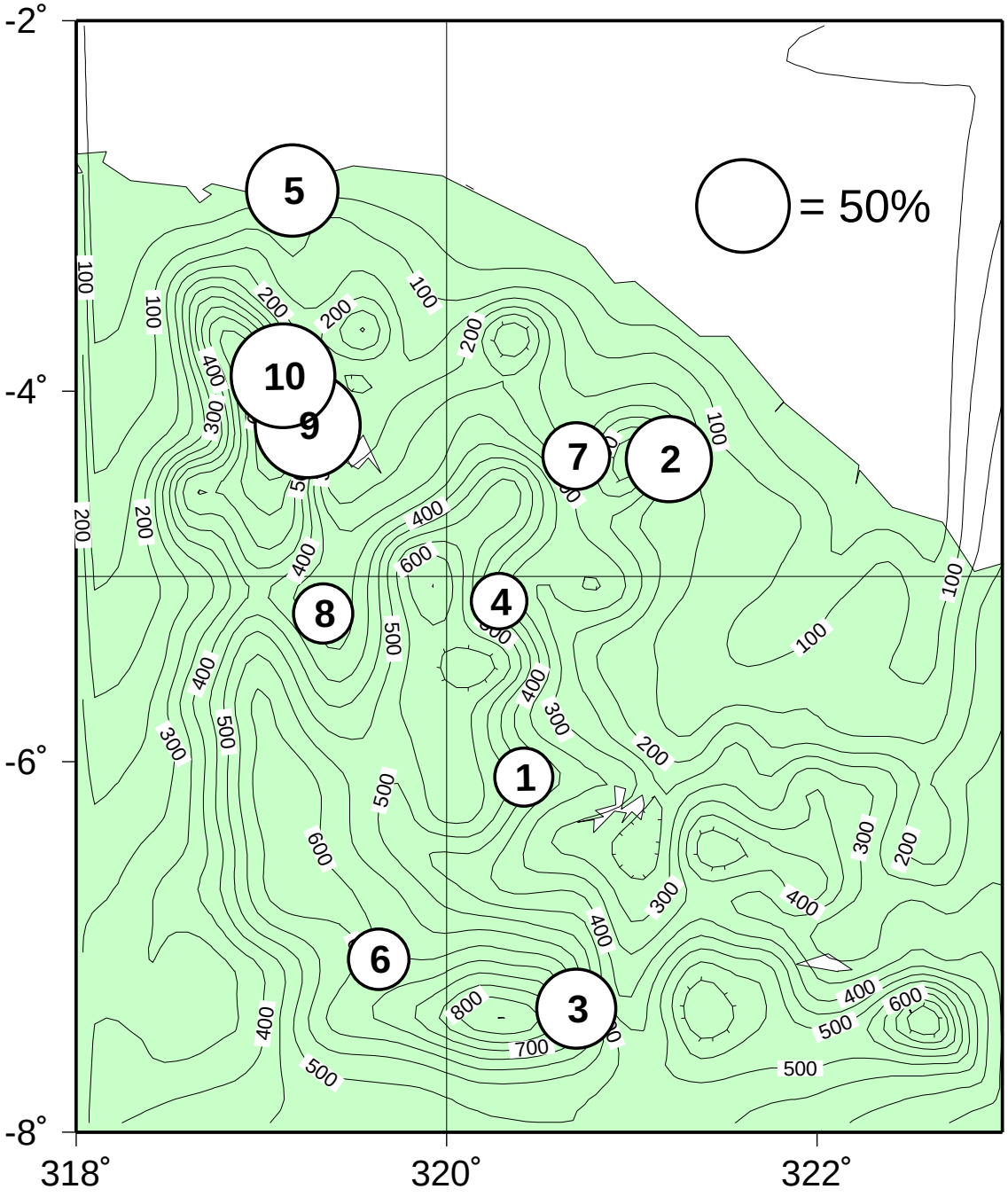

Figure 1: Rainfall station locations with topographic contours (meters). Circle size denotes the February–April climatological daily rainfall probability (%) 1975–2002. The stations are: (1) Acopiara (317 m), (2) Aracoiaba (107 m), (3) Barbalha (405 m), (4) Boa Viagem (276 m), (5) Camocim (5 m), (6) Campos Sales (551 m), (7) Caninde (15 m), (8) Crateus (275 m), (9) Guaraciaba Do Norte (902 m), and (10) Ibiapina (878 m). One degree of longitude/latitude corresponds to about 110 km at the equator.

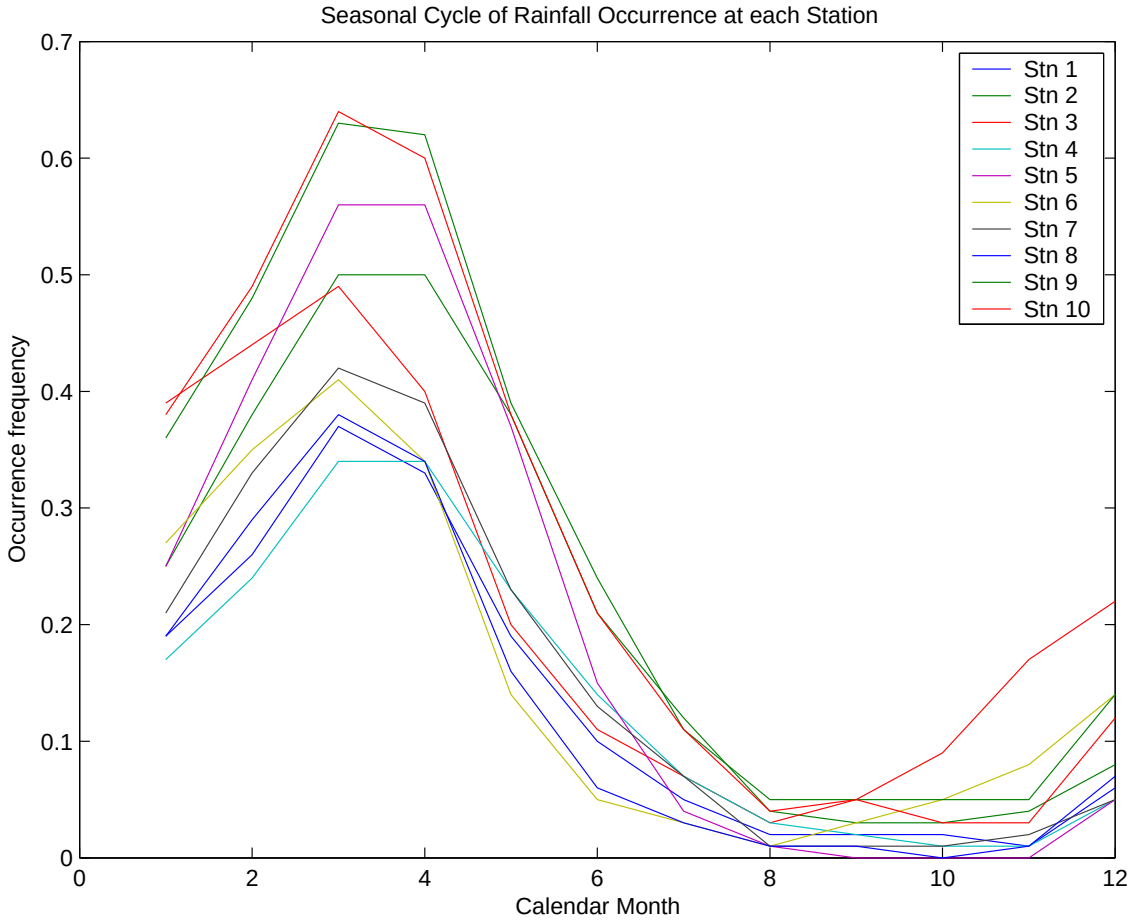

Figure 2: The mean seasonal cycle of rainfall occurrence frequency at each station over Ceará.

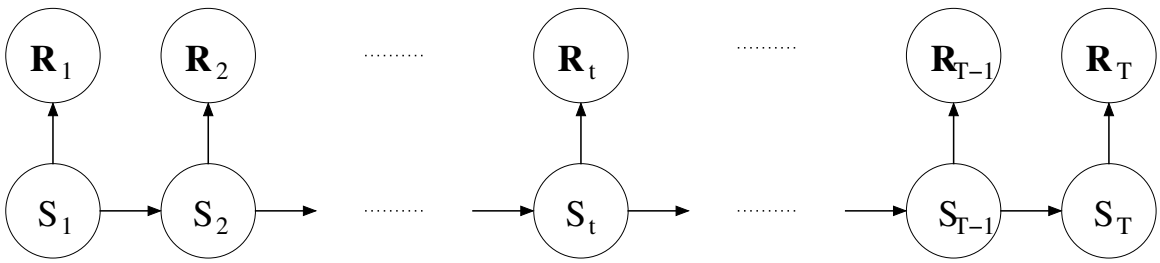

Figure 3: Graphical model representation of a hidden Markov model.

The second assumption is that the hidden state process, $S_{1:T}$, is first-order Markov:

$$P\left(S_t|S_{1:t-1}\right) = P\left(S_t|S_{t-1}\right) \tag{2}$$

and that this first-order Markov process is homogeneous in time, i.e., the $K \times K$ transition probability matrix for Equation 2 does not change with time.

For $P\left(\mathbf{R}_t|S_t\right)$, we make the simplifying assumption that the rainfall observation at each station at time t is independent from observations at other stations at time t, conditional on the hidden state:

$$P\left(\mathbf{R}_t = \mathbf{r}|S_t = s\right) = \prod_{m=1}^{M} P\left(R_t^m = r|S_t = s\right) = \prod_{m=1}^{M} p_{smr}$$

where $r \in \{0, 1\}$, each $p_{smr} \in [0, 1]$, and $p_{sm0} + p_{sm1} = 1$. Note that this *conditional* independence assumption given the states does not imply spatial independence of the rainfall process (which would be unreasonable). Spatial dependence is captured implicitly via the state variable in the model, as we will see later in the paper.

Techniques for parameter fitting and prediction using homogeneous HMMs are well known in the statistical literature and, thus, need not be elaborated on here—the reader is referred to standard references such as Rabiner (1989) and MacDonald and Zucchini (1997) for details. Figure 3 shows a graphical representation of the conditional independence assumptions in the HMM. Later in the paper we develop a non-homogeneous extension of this model where the transition probabilities are no longer homogeneous in time but instead are allowed to vary over time.

4 Hidden States of Daily Rainfall Occurrence

In this section, we use the homogeneous HMM to construct states of daily rainfall occurrence from the 10-gauge daily record. As a baseline we also evaluate the performance of a model with no hidden states, where independent Markov chains were fit to the historical data from each station using maximum likelihood. The stateless model is equivalent to the classical "weather generator," commonly used to used to model daily rainfall occurrence (e.g., Wilks and Wilby 1999).

We used cross-validation to evaluate the quality of the fitted HMMs as a function of K the number of states. HMMs for different values of K were fit to the training data leaving out different one-fourths (six consecutive years) of the data at a time and then calculating the log-probability of the observed data for the left-out years. The resulting normalized out-of-sample values of the

Table 1: Performance of HMMs with different numbers of states K compared to a model with independent Markov chains.

Model	Out-of-Sample Normalized Cross-Validated Log-Likelihood	Normalized BIC Score	Out-of-Sample Average Absolute Difference in Spatial Correlation (per station pair)	Out-of-Sample Average Absolute Difference in Rainfall Persistence
Markov Chains	0.6314	0.6325	0.2453	0.0641
$K = 2$	0.5856	0.5844	0.0627	0.0793
$K = 3$	0.5743	0.5743	0.0511	0.0678
$K = 4$	0.5709	0.5723	0.0465	0.0635
$K = 5$	0.5687	0.5725	0.0460	0.0644
$K = 6$	0.5684	0.5737	0.0450	0.0651

cross-validated log-likelihood for each model are given in Table 1 for $K = 2, 3, 4, 5, 6$, together with a normalized Bayes Information Criterion (BIC) for each model (see Appendix B).

The *normalized* cross-validated log-likelihood is defined to be the negative of the original total (not averaged over years) cross-validated log-likelihood divided by the total number N of binary events used in evaluating the model. Here $N = 24 \times 90 \times 10$, the number of individual rainfall events across all years, days, and stations. Scaled in this fashion the normalized log-likelihood corresponds to a form of predictive entropy or uncertainty of the model in bits (for base2 logarithms). The resulting normalized scores are somewhat more interpretable than unnormalized log-likelihood scores since they lie on a scale between 0 and 1. The BIC scores in Table 1 are scaled in a similar manner (Appendix B).

Not surprisingly in Table 1 the HMMs have lower predictive uncertainty and lower normalized BIC scores than the independent chains model (where lower scores correspond to better predictive ability). The cross-validated out-of-sample log-likelihood decreases substantially from $K = 2$ to $K = 4$, and then levels off. The normalized BIC score reaches its minimum at $K = 4$. Given that both scores indicate that $K = 4$ is a reasonable choice for K, this is the value we chose in the following.

Measures of the models' ability to reproduce both (a) observed spatial correlations between stations and (b) rainfall persistence (the probability of rain given rain the previous day for a particular station) are also included in the table. Each number corresponds to the average absolute difference between (a) 3000 years of simulated data from a model trained on 18 years of data, and (b) actual data from the remaining 6 out-of-sample years. In this manner the numbers in Table 1 indicate the out-of-sample predictive power of the model in terms of both spatial correlation and rainfall persistence. The baseline average absolute error for a model which has no spatial correlation is 0.2453. This is also the error of the independent chains model since this model has (by definition) zero spatial correlation in its simulations (over the long run). All of the HMMs reduce this error by roughly a factor of 4, or equivalently can model roughly 80% of the total daily spatial correlation in the data. The HMM models achieve this via the hidden states which can implicitly capture

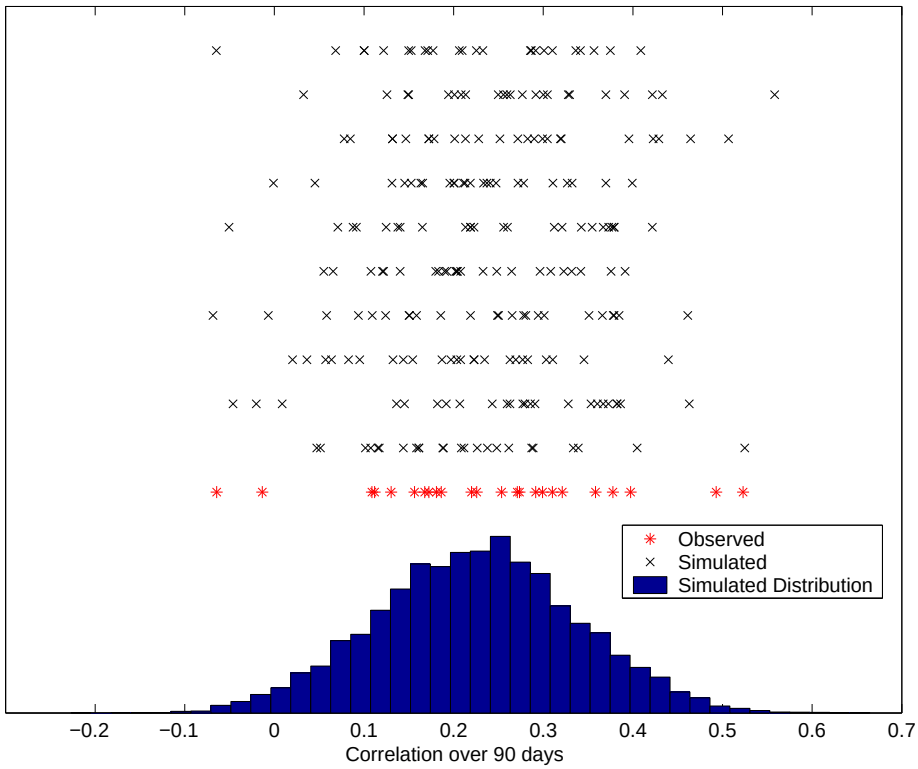

Figure 4: Observed correlations between stations 1 and 8 for all 24 years of data compared with correlations between the same stations from the data simulated 10 times (plotted stacked above each other) from the 4-state HMM trained on all 24 years. The histogram of the simulated distribution was computed from 12000 90-day seasons.

marginal spatial dependence.

There is less difference between the HMMs and the independent chains model in terms of persistence. We might expect the independent Markov chains model to outperform any HMM since wet and dry spells in daily rainfall data are often well-modeled as first-order Markov (Wilks and Wilby 1999). Empirically, in Table 1 the HMMs provide similar persistence numbers to that of the independent chains, out-of-sample. The specific value of K for the HMMs does not seem to have a large impact on the spatial correlation or persistence error numbers, apart from a slight flattening out in error after $K = 4$.

The correlation and persistence numbers in Table 1 summarize the difference in expected values for the statistics over multiple years. We find that the HMM trained on the data can also recover most if not all of the year-to-year variability of the statistics. Figures 4 and 5 show for a selected pair of stations (for correlation) and for one station (for persistence) the observed 24 yearly values of the statistic. Also shown for comparison are 10 runs producing 24 yearly values of the same statistic produced from an HMM trained on the observed data. Statistics from the simulated data appear to be similar to the ones calculated from the observed data. Moreover, the variability in statistics for the observed years generally agrees with the empirical probability distribution of simulated statistics (using 12000 sequences or years).

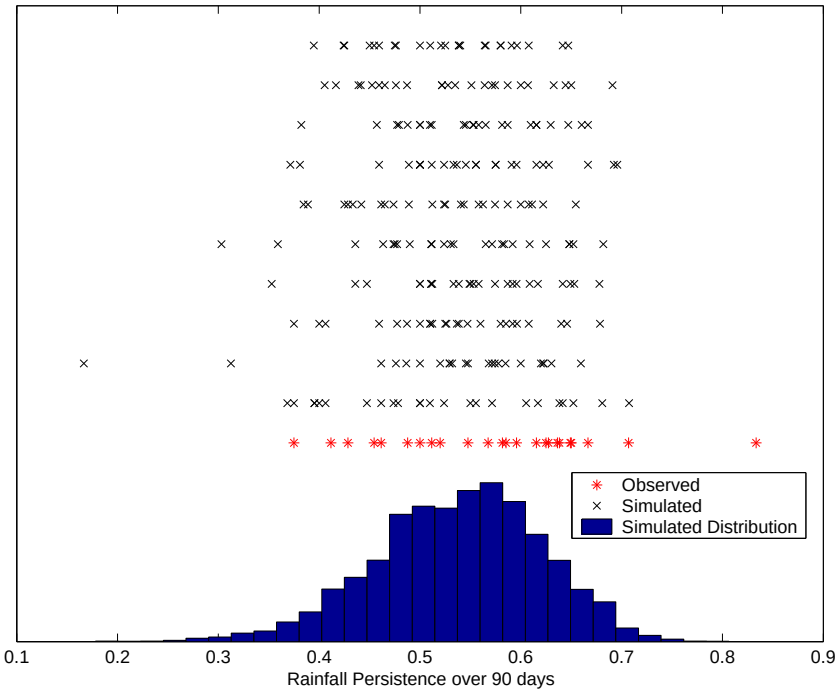

Figure 5: Observed persistence for station 5 for all 24 years of data compared with persistence for the same station from the data simulated from the 4-state HMM trained on all 24 years. Details of simulations as in Fig. 4.

In general, the distributions of wet and dry spell-lengths are captured quite well by the HMM, as illustrated in Fig. 6. The spell-length distributions (both simulated and observed) follow approximately geometric distributions (near-straight lines in the semi-log plot on Fig. 6), while the HMM tends to underestimate the spell durations at some stations. The geometric distribution is a characteristic of the Markov model. However, the HMM treats the state sequence as a Markov chain, rather than the rainfall. Thus, while the states in data simulated from an HMM will have run-lengths whose distribution is geometric, the observed precipitation may be more "bursty", leading (for example) to possible underestimation of rainfall persistence.

The state transition-probability matrix is given in Table 2. The self-transitions are relatively large indicating that the states are persistent, with states 1 and 2 being the most so, and state 4 being the least persistent. Direct transitions between states 1 and 2 are rare, with states 3 and (especially) 4 playing the role of intermediaries. There are no very clear transition directions, though state 1 tends to follow state 4 ($p = 0.22$) rather than precede it ($p = 0.11$).

The rainfall probabilities for each state are plotted in Fig. 7, along with the number of days assigned to each state. The four states fall roughly into two pairs with states 1–2 characterized by wet or dry conditions at all stations and states 3–4 describing anomalous north-south gradients (see also Fig. 9 which shows the probabilities as anomalies from climatology). State 1 (state 2) is characterized by increased (decreased) probability of rain at all stations, compared to the climatological probabilities in Fig. 1; state 2 has near-zero probability of rain everywhere except on the coast. State 3 (state 4) has anomalously small (large) probabilities in the south and slightly

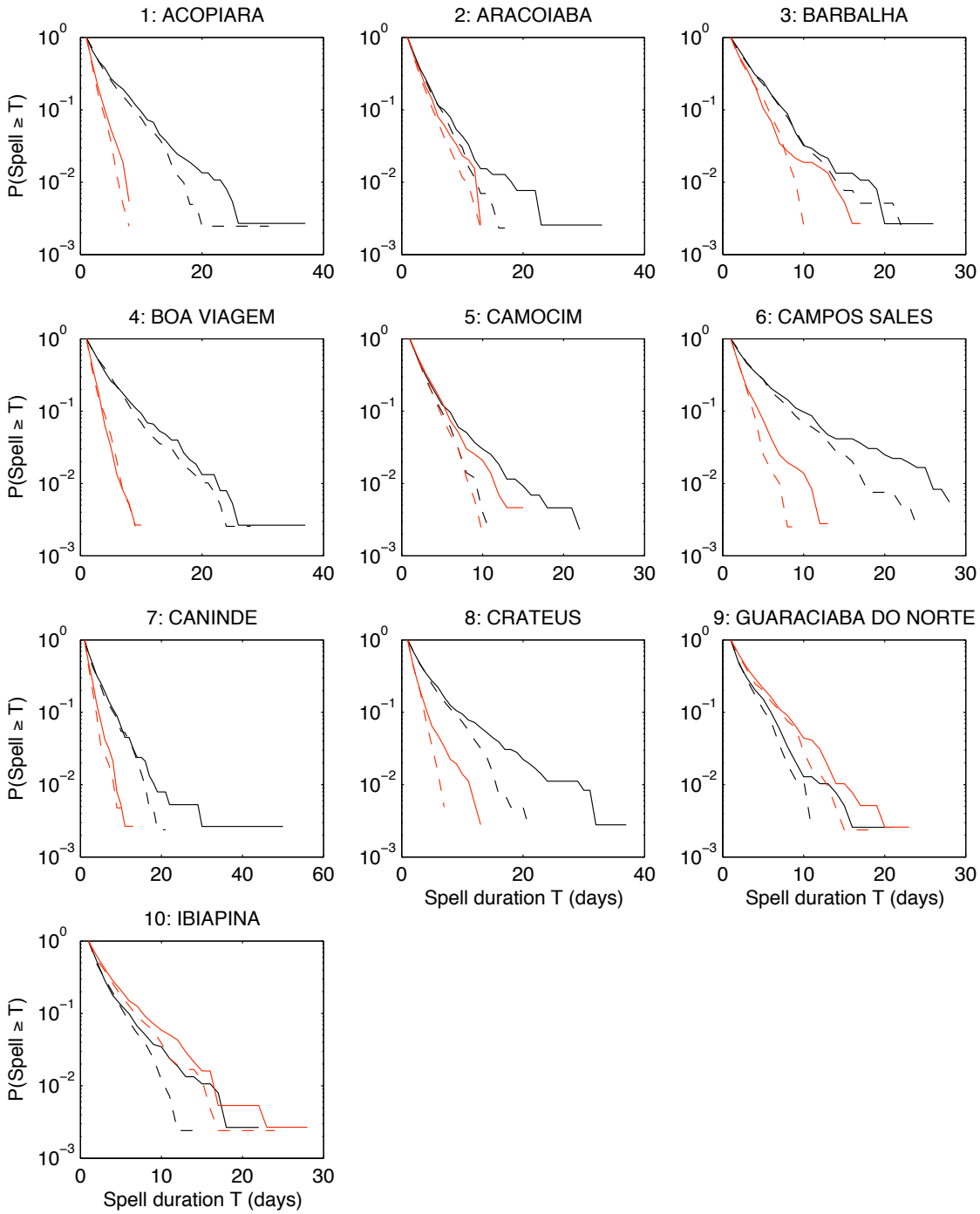

Figure 6: Distribution of spell-lengths for wet (red) and dry (thick) spells at each station. Dashed lines are computed from the 4-state HMM simulations, with the observed data shown as solid lines. Plotted is the probability of a spell exceeding a particular duration. A geometric distribution would plot as a straight line on these semi-log plots.

Table 2: Transition probabilities for 4-state HMM.

		to state			
		1	2	3	4
	1	0.70	0.01	0.18	0.11
from	2	0.02	0.68	0.16	0.13
state	3	0.18	0.14	0.61	0.08
	4	0.20	0.20	0.12	0.48

increased (reduced) probabilities in the north, with near-climatological probabilities in the center of the domain. It is notable that the rainfall probabilities for states 1 and 2 are more spatially uniform than in the climatology, while the opposite is true of states 3 and 4 which are characterized by larger meridional gradients.

The estimated state-sequence, derived using the Viterbi algorithm (Viterbi 1967), is shown in Fig. 8. The sequence exhibits considerable variability on intra-seasonal, as well as interannual time scales. We now examine systematic seasonal and interannual variability.

The mean seasonality of state-occurrence is shown in Fig. 10. The frequency of state 2 decreases from early February to mid-March, while the prevalence of state 1 maximizes in March, indicating the peak of the wet season at all stations. State 4 decreases in prevalence toward the end of the rainy season, while state 3 tends to become more frequent, indicating a contraction of the wet season northward. The ratios of minimum/maximum frequencies (from 24-year averages of 10-day running means) plotted in Fig. 10 are 0.32, 0.25, 0.36 and 0.28 for states 1–4 respectively; i.e. the state-frequency varies by a factor of 3–4 within the season.

We now examine the meteorological characteristics associated with each rainfall state, by compositing anomalies of NOAA interpolated outgoing longwave radiation (OLR) and NCEP/NCAR reanalysis winds over the days assigned to each state. Figure 9 shows anomaly-composites of OLR and 850-hPa winds for the tropical South American-Atlantic sector, along with the state rainfall probabilities displayed as anomalies from the FMA-mean climatology. Note that the anomalies in winds and OLR are computed here as deviations from the mean seasonal cycle. States 1 and 2 are clearly associated with strongly contrasting large-scale anomalies in OLR and the cross-equatorial flow. Together these represent north-south displacements in the Atlantic ITCZ, as well as zonal contractions or expansions of the SAMS core region over Amazonia.

State 1 represents an anomalously southward-displaced ITCZ, together with an eastward expansion of SAMS. Thus, the ITCZ and SAMS merge to a greater extent than in the FMA climatology, consistent with state 1 tending to correspond to the peak of the seasonal evolution (Fig. 10). In this configuration, Ceará comes entirely under the influence of the large-scale convection zones, and rainfall probabilities are high at all 10 stations. The contrasting situation characterizes state 2, in which SAMS and the ITCZ become more separated, reminiscent of the DJF climatology, with Ceará largely dry. It is worth emphasizing that states 1 and 2 are associated with sizeable atmospheric anomalies with respect to the seasonal cycle (i.e. those plotted in Fig. 9), so that these states are not merely aspects of the latter. Indeed the anomalous winds associated with states 1

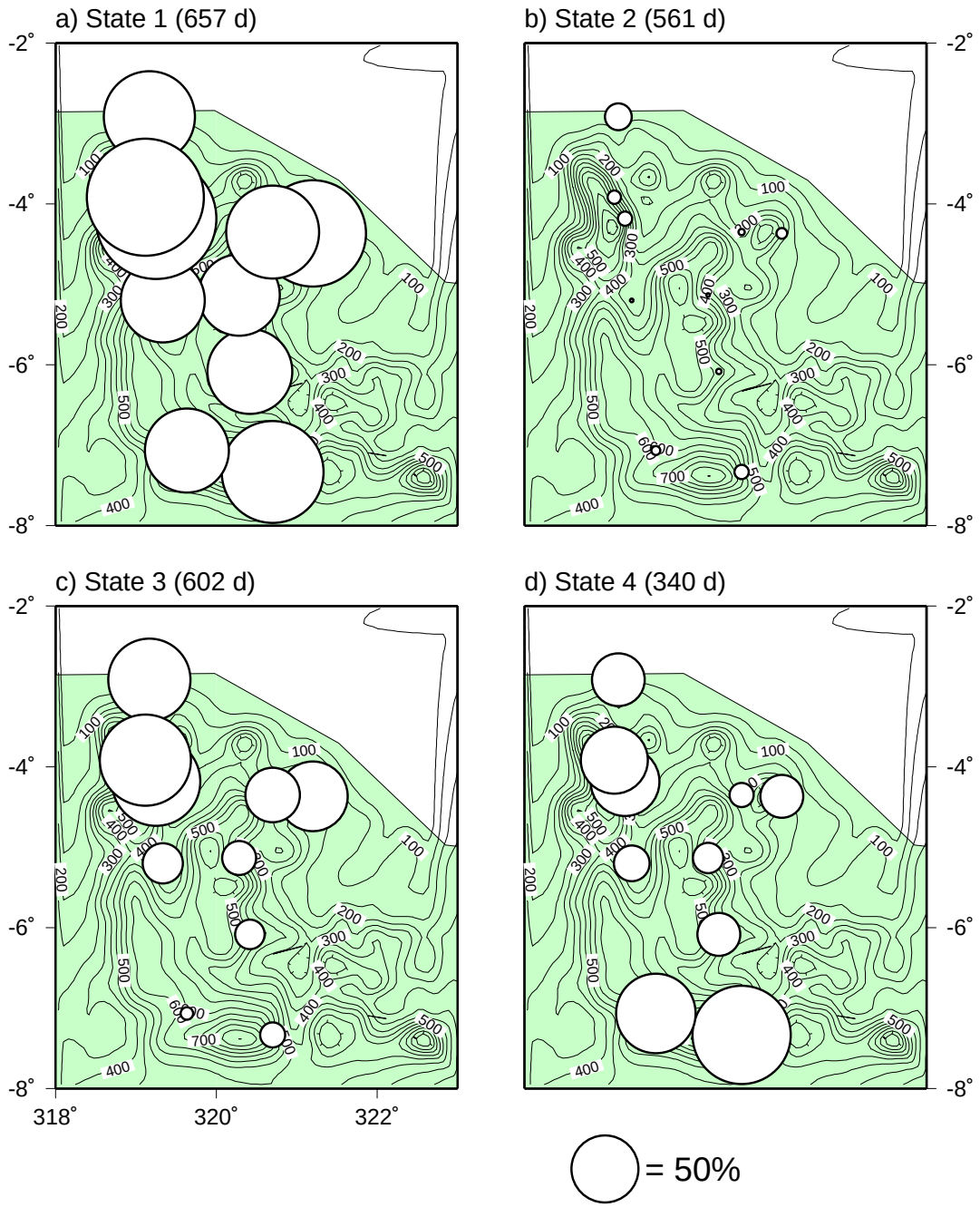

Figure 7: Four-state HMM rainfall probabilities (%) (circle size) together with topographic contours. The number of days in each state is given in brackets.

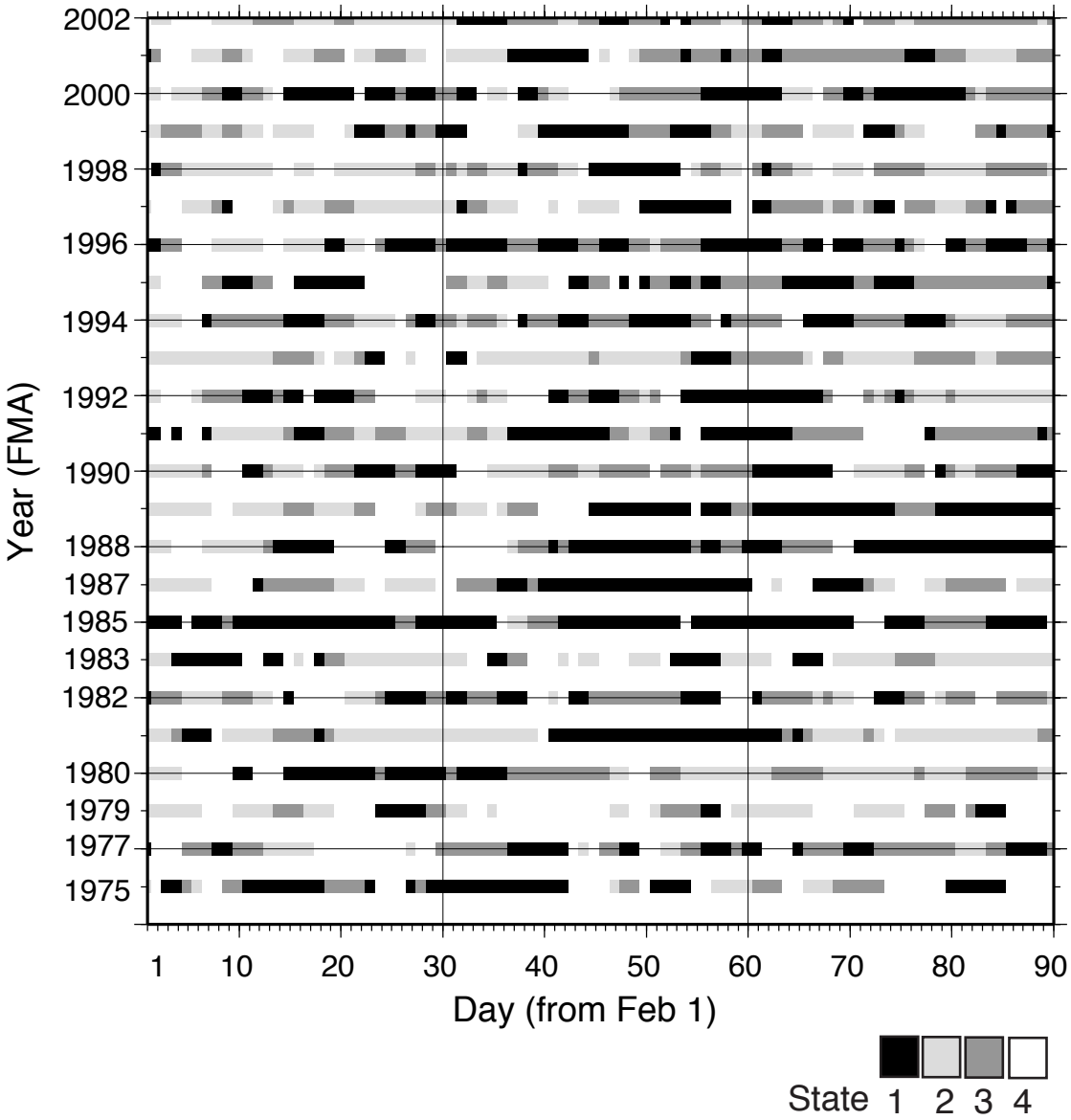

Figure 8: The estimated state sequence.

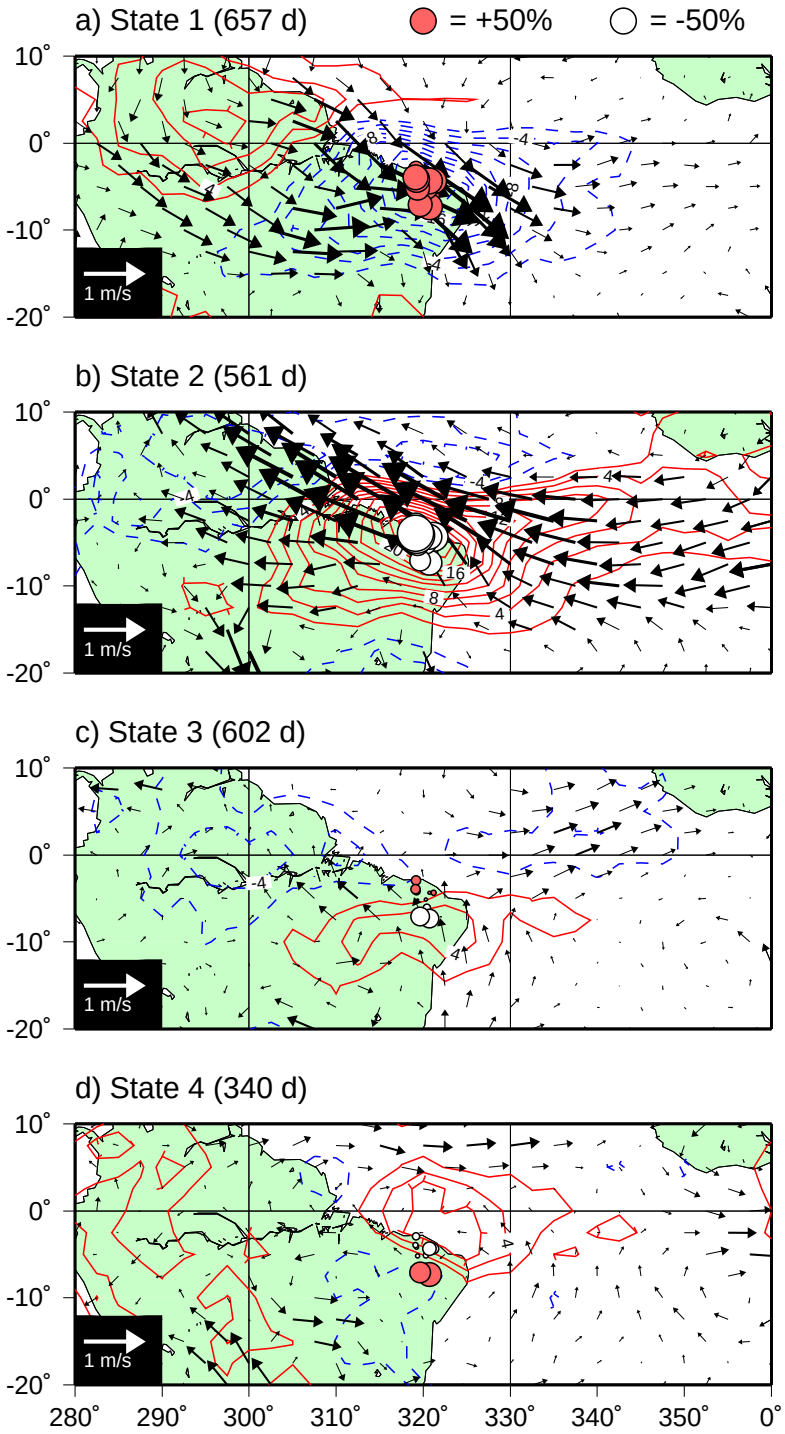

Figure 9: HMM state anomaly composites of 850-hPa wind (vectors) and OLR (contours), together with rainfall anomaly-probabilities (%) at each station (circles). The wind and OLR anomalies are deviations from the mean seasonal cycle; the latter was computed by 10-day lowpass filtering and then averaging all years together. The composites are defined on the days assigned to each state, with the number of days given in the caption. Contour interval: 2 W m^{-2}, with negative contours dashed.

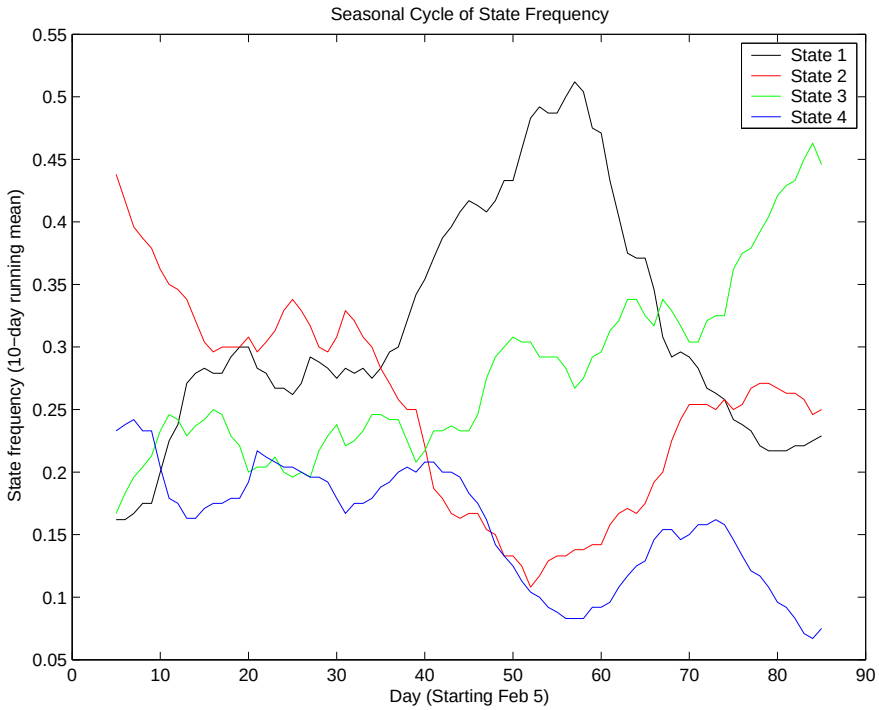

Figure 10: Mean seasonal variation of four HMM daily state occurrence, smoothed with a 10-day running mean prior to averaging.

and 2, plotted in Fig. 9, change little if anomalies are computed with respect to the FMA long-term mean.

States 3 and 4 are associated with smaller anomalies in rainfall probability, and this is reflected in much smaller anomalies in OLR and wind. The north-south gradients in the OLR anomalies are consistent with those in the gauge-rainfall HMM probabilities. However, the wind anomalies are not very coherent.

In order to identify any large-scale atmospheric teleconnections, Fig. 11 shows composites of 850- and 200-hPa wind anomalies of states 1 and 2 over a larger domain. Over NE Brazil, the direction of the wind anomalies reverses with height, typical of the first baroclinic mode of tropical atmospheric dynamics. Over the tropical Pacific, states 1 and 2 are characterized by anomalous Walker Circulations, and these are consistent with OLR anomalies over the equatorial Pacific (not shown). Over the Atlantic, they suggest opposite polarities of the Atlantic meridional mode, with NAO-like wind anomalies and anomalies in the NE Trades. These are recognized to be the two mechanisms that influence NE Brazil rainfall on interannual time scales, and both leave their imprint on the daily rainfall states.

Similar pictures for states 3 and 4 (not shown) do not highlight any coherent circulation patterns that might be remotely forcing these rainfall states. However, there is a hint, especially for state 3, that Rossby wave activity propagating from the South Pacific may be a contributing factor (cf. Liebmann et al. 1999).

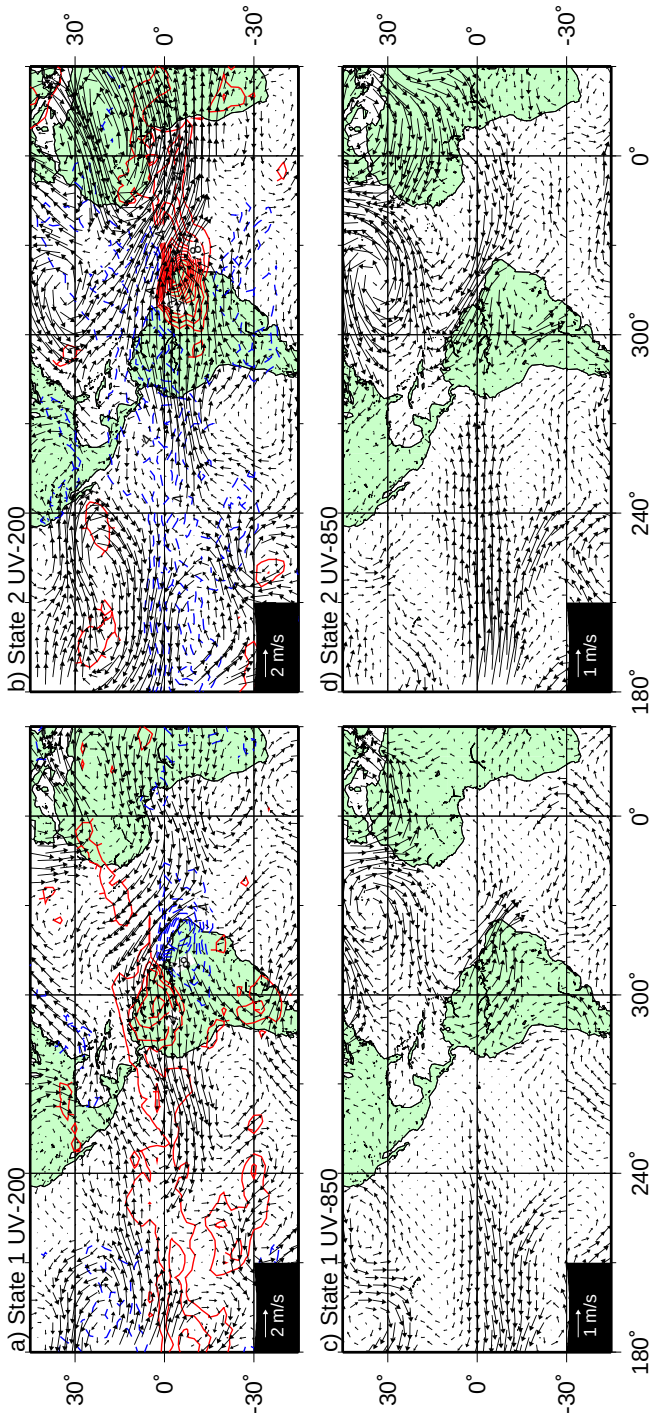

Figure 11: HMM state anomaly composites of wind over a larger domain for states 1 and 2. (a–b): 200-hPa winds. (c–d): 850-hPa winds. Details as in Fig. 9.

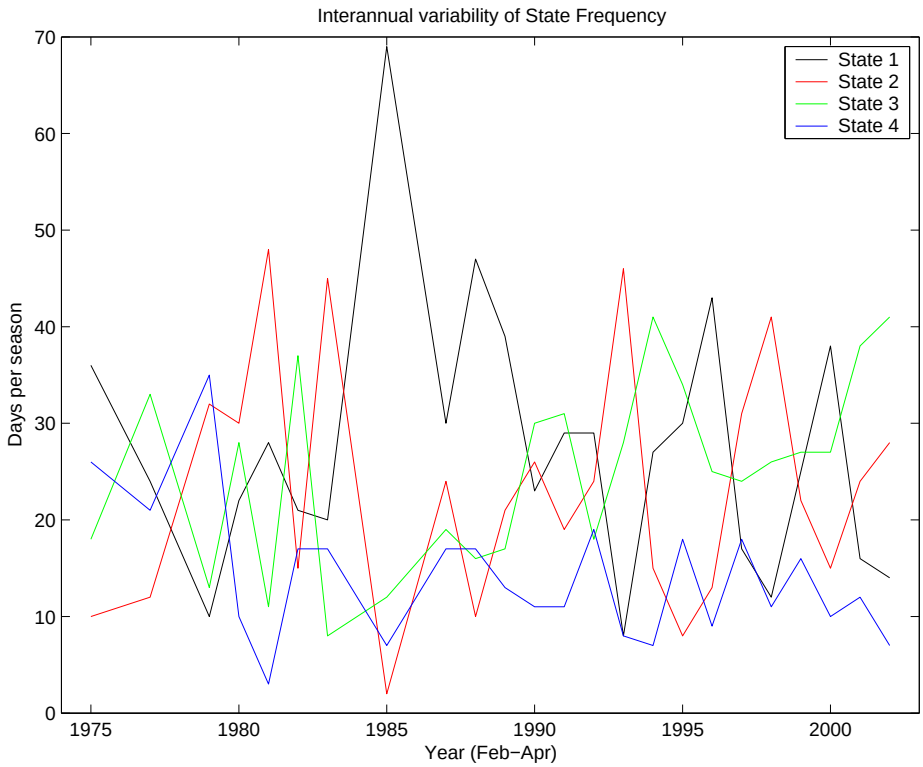

Figure 12: Interannual variability of HMM state-occurrence frequency.

Interannual variability of state-occurrence is plotted in Fig. 12 in terms of the number of days assigned into each state. Large interannual variations do occur, especially in states 1 and 2, which vary in opposition to each other consistent with their rainfall and meteorological characteristics seen in Figs. 9 and 11. The occurrence frequency of state 3 shows an upward trend from the 1980s onwards, while state 4 shows little interannual variation.

To probe the interannual variability further, composites of seasonal-mean SST anomalies are shown in Fig. 13 for FMA seasons in which the interannual state-frequency (Fig. 12) is in the top 15% of years (i.e., 4 or 5 years). The shading depicts significance at the 90% level. A clear ENSO SST anomaly signature is present for years in which states 1 or 2 are highly prevalent, but statistical significance is only high for state 1 (La Niña). States 1 and 2 are also associated with Atlantic SST anomalies characteristic of the Atlantic meridional mode, but these are weak. States 3 and 4 are not associated with appreciable SST anomalies.

Finally, to check whether these four specific states might be sensitive to our choice of model (i.e., the HMM), we also ran the well-known K-means clustering algorithm (e.g., Jain and Dubes 1988) on the data with $K = 4$ clusters. Here the input was a set of 10-dimensional vectors with binary components corresponding to the daily rainfall occurrence measurements. The four K-means clusters were found to match closely those derived from the HMM. Specifically, the 10-dimensional means for each cluster from K-means (real-valued numbers between 0 and 1) were compared to the conditional probability vectors for each state from the HMM, and the composite wind and OLR maps resulting from the most likely assignments of each day to each cluster (for both K-means and

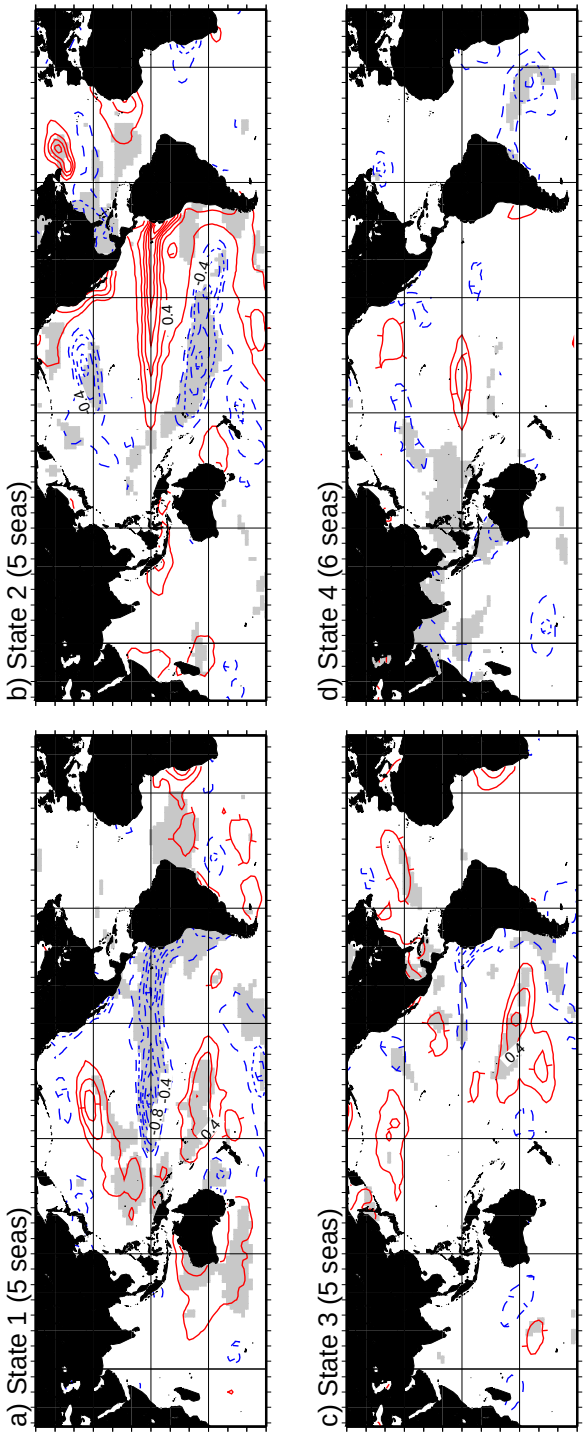

Figure 13: Composites of anomalous SST for years (February–April) when HMM states are most prevalent, defined as years in the upper 15% of the interannual distribution of state frequency. The number of years in each composite is given in brackets. Shading represents 90% statistical significance according to a two-sided Student t-test. Negative contours are dashed, and the zero contour is omitted. Contour interval: 0.2^oC.

HMMs) were also visually compared. The means and maps obtained from the two different methods were found to be quite similar (not shown). The fact that an alternative clustering methodology such as K-means, that uses no information about temporal ordering of the rainfall measurements, produces state descriptions that are qualitatively similar to those produced by the HMM, suggests that these states are an inherent property of the data and insensitive to the particular modeling methodology being used.

In summary, daily rainfall states 1 and 2 identified by the HMM are associated with well-known patterns of interannual variability in winds, OLR and SST. These associations provide a basis for the downscaling of seasonal GCM predictions, and this is pursued in the following section.

5 A NonHomogeneous HMM Downscaling Prototype

The NHMM generalizes the homogeneous HMM in that the transition probabilities in Equation 2 are allowed to vary with time. In particular, for downscaling applications the transition probabilities between states are allowed to vary as a function of external inputs. Hughes and Guttorp (1994) introduced this model in the context of modelling rainfall occurrence. The NHMM used in this paper is based on this original work of Hughes and Guttorp, with some minor modifications.

In this section we illustrate the ability of an NHMM to downscale atmospheric GCM simulations over NE Brazil. It is found that introducing atmospheric input variables does not visibly change the appearance of the state composites, nor appreciably change the rainfall probabilities. Thus, a 4-state model is chosen for consistency with the HMM in the previous section.

For demonstration purposes and for consistency with IRI's current seasonal-forecast scheme, we define the inputs to the NHMM from the GCM's simulated seasonal-mean rainfall anomaly. The daily values needed as inputs to the NHMM are derived by simply repeating the seasonal-mean input value for each day within the FMA season.

5.1 The NonHomogeneous Hidden Markov Model

Let $\mathbf{X}_t$ be a D-dimensional column vector of predictors for day t, derived for example from a GCM. By $\mathbf{X}_{1:T}$ we will denote the sequence $\mathbf{X}_1, \ldots, \mathbf{X}_T$. We now replace Equation 2 in the homogeneous HMM with:

$$P\left(S_t | S_{1:t-1}, \mathbf{X}_{1:T}\right) = P\left(S_t | S_{t-1}, \mathbf{X}_t\right), \tag{3}$$

so that the hidden state on day t depends both on the predictor vector $\mathbf{X}_t$ for day t and the value of the hidden rainfall state S_{t-1} on day $t-1$. Because $\mathbf{X}_t$ can vary in time, this results in transition probabilities between states that are can vary in time in response to changes in $\mathbf{X}$, i.e., an inhomogeneous model. The corresponding graphical model is shown in Fig. 14.

The hidden state transitions in Equation 3 are modelled by a polytomous (or multinomial) logistic regression:

$$P\left(S_t = i | S_{t-1} = j, \mathbf{X}_t = \mathbf{x}\right) = \frac{\exp\left(\sigma_{ji} + \boldsymbol{\rho}_i' \mathbf{x}\right)}{\sum_{k=1}^{K} \exp\left(\sigma_{jk} + \boldsymbol{\rho}_k' \mathbf{x}\right)}. \tag{4}$$

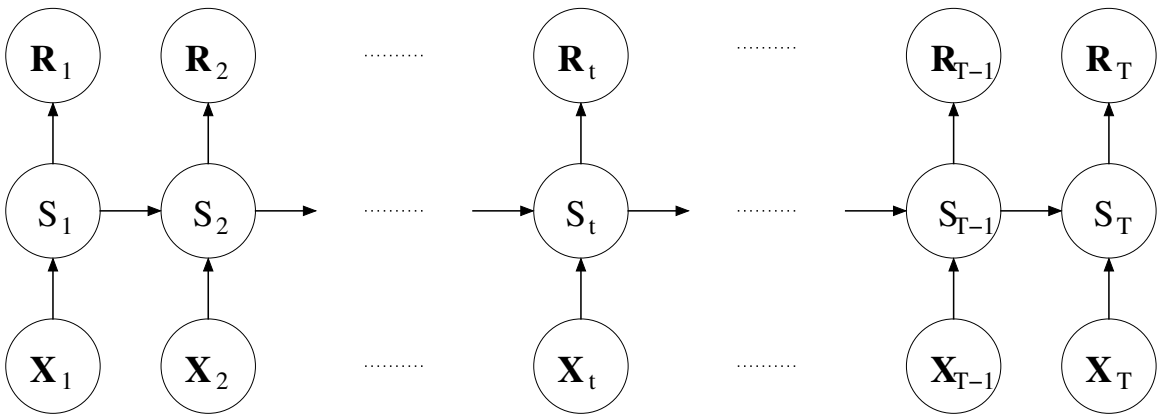

Figure 14: Graphical model representation of a non-homogeneous hidden Markov model.

For the specific case of S_1, the first hidden state in the sequence,

$$P\left(S_1 = i | \mathbf{X}_1 = \mathbf{x}\right) = \frac{\exp\left(\lambda_i + \boldsymbol{\rho}_i' \mathbf{x}\right)}{\sum_{k=1}^{K} \exp\left(\lambda_k + \boldsymbol{\rho}_k' \mathbf{x}\right)}. \tag{5}$$

All σ's and λ's are real-valued parameters while the $\boldsymbol{\rho}$'s are D-dimensional real-valued parameter vectors, where the prime denotes the vector transpose. This parameterization can be shown to be equivalent to the one defined in Hughes and Guttorp (1994) and in Hughes et al. (1999) in which the base-line transition matrix is multiplied by a function of the atmospheric predictors:

$$\begin{aligned} P\left(S_t = i | S_{t-1} = j, \mathbf{X}_t\right) &\propto P\left(S_t = i | S_{t-1} = j\right) P\left(\mathbf{X}_t | S_{t-1} = j, S_t = i\right) \\ &= \gamma_{ji} \exp\left(-\frac{1}{2}\left(\mathbf{X}_t - \boldsymbol{\mu}_{ji}\right)' \mathbf{V}^{-1}\left(\mathbf{X}_t - \boldsymbol{\mu}_{ji}\right)\right) \\ &\propto \exp\left(\left(\ln \gamma_{ji} - \frac{1}{2}\boldsymbol{\mu}_{ji}' \mathbf{V}^{-1} \boldsymbol{\mu}_{ji}\right) + \boldsymbol{\mu}_{ji}' \mathbf{V}^{-1} \mathbf{X}_t\right). \end{aligned} \tag{6}$$

Here $\boldsymbol{\mu}_{ji}$ is the mean of the atmospheric predictor-vector associated with transitions from state j at day $t-1$ to state i at day t. If we make the simplification that $P\left(\mathbf{X}_t | S_t, S_{t-1}\right) = P\left(\mathbf{X}_t | S_t\right)$, then $\boldsymbol{\mu}_{ji} = \boldsymbol{\mu}_i$, and the parameterization in Equation 6 becomes equivalent to the one in Equation 4. This can be shown by setting $\lambda_i = \ln \gamma_{ji} - \frac{1}{2}\boldsymbol{\mu}_{ji}' \mathbf{V}^{-1} \boldsymbol{\mu}_{ji}$ and $\rho_i = \mathbf{V}^{-1} \boldsymbol{\mu}_{ji}$.

The parameters λ_1, σ_{j1}, and $\boldsymbol{\rho}_1$ are set to zero to guarantee the identifiability of the transition parameters. Note that the homogeneous transition matrix (Equation 2) can be viewed as a special case where $\boldsymbol{\rho}_i = \mathbf{0}$ for all $i = 1, \ldots, K$.

An example of the transition probabilities obtained with this parameterization for a 4-state model with a univariate normalized input is shown in Fig. 15. This case corresponds approximately to the downscaling example to be presented below, and demonstrates how the transition probabilities from state 2 to states 1–4 are modulated by the value of X_t. Each curve can be intrepreted as part of a logistic "S-shaped" curve, with the value of $X = 0$ corresponding to the homogenous HMM, and the central portion of the plot being the most relevant.

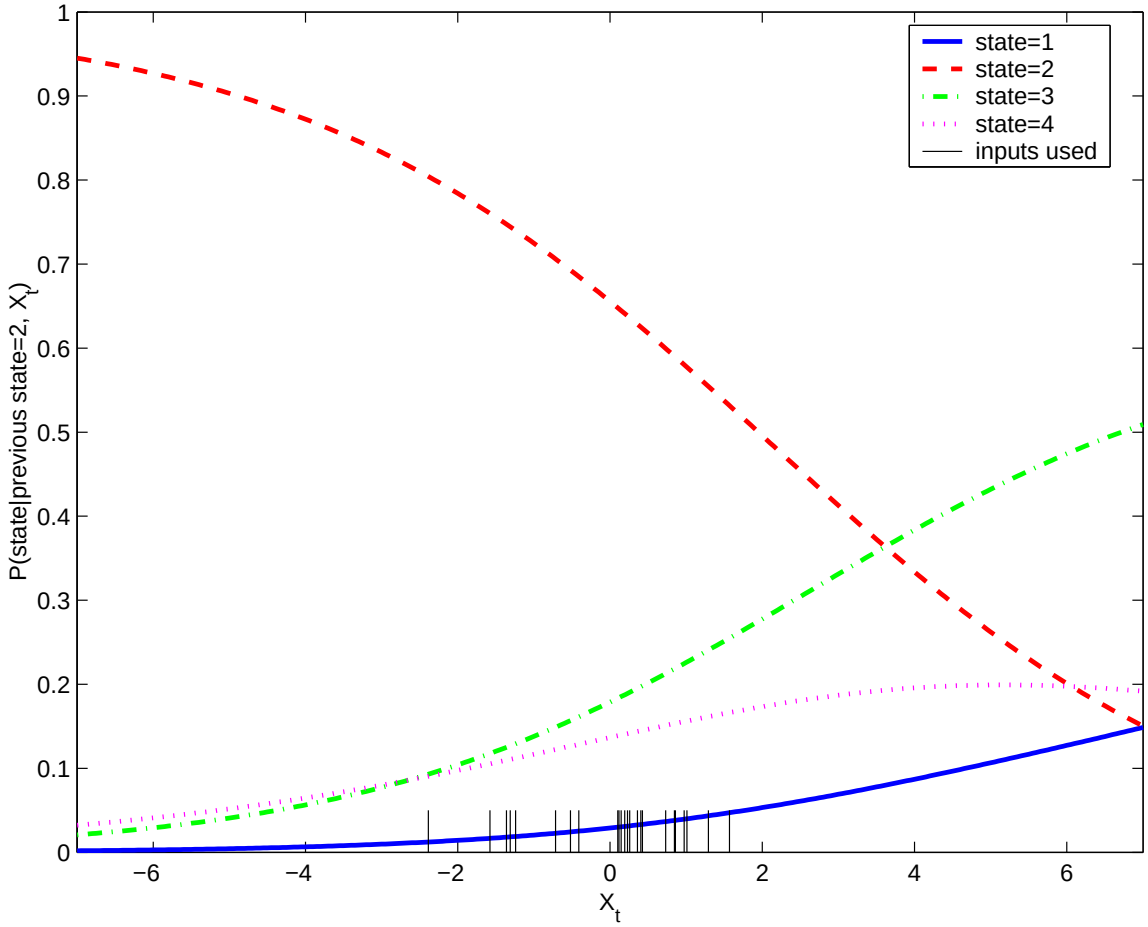

Figure 15: Transition probabilities into each state from state 2 ("dry"), as a function of a univariate input variable for a 4-state NHMM. The input variable is defined by the ensemble-average GCM simulation of seasonal-average precipitation, averaged spatially over the region of NE Brazil (centered and normalized by its standard deviation). The values of X_t used to train the model are plotted along the abscissa.

Given a fixed number of hidden states K, we learn the parameters $\mathbf{\Theta}$ of the NHMM by searching for parameters that best fit the observed data. To do this, we employ the commonly used maximum likelihood principle. Specifically, we search for $\mathbf{\Theta}$ that maximizes the conditional probability of the observed data as a function of $\mathbf{\Theta}$—this conditional probability function is referred to as the likelihood:

$$L(\mathbf{\Theta}) = P(\mathbf{R}_{1:T}|\mathbf{X}_{1:T}, \mathbf{\Theta}) = \sum_{S_{1:T}} P(S_1|\mathbf{X}_1) \prod_{t=2}^{T} P(S_t|S_{t-1}, \mathbf{X}_t) \prod_{t=1}^{T} P(\mathbf{R}_t|S_t). \tag{7}$$

The set of parameters $\mathbf{\Theta}$ that maximize $L(\mathbf{\Theta})$ can be obtained using the well-known Baum-Welch algorithm (e.g., Rabiner 1989), a variation of an iterative Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) for obtaining maximum likelihood parameter estimates for models with hidden variables and/or missing data. Full details on the specific EM procedure used in this paper for NHMM parameter estimation can be found in the Appendix A.

5.2 Predictor selection: Canonical Correlation Analysis

The choice of input ("predictor") variables is non-trivial. Here, canonical correlation analysis (CCA) between historical seasonally-averaged rainfall occurrence probabilities at each of the 10 stations, and the GCM's simulated seasonal-mean rainfall amount at gridpoints within the region $(80^oW - 0^oW, 21^oS - 7^oN)$ was used to define the input variables. The CCA yields a calibration of the GCM's simulation of rainfall amount with respect to observed rainfall occurrence probabilities. The 24-member ensemble mean of the GCM's normalized precipitation field was used, driven by historical estimates of SSTs. The analysis proceeds by firstly expanding each field into principal components (PCs), and then performing the CCA in the reduced subspace of the two sets of PCs. The GCM precipitation was normalized by its local interannual standard deviation prior to the CCA. The optimal truncations for each PC subspace, as well as the optimal number of CCA modes were determined by (1) computing the CCA for a given choice of truncation and, (2) summing the out-of-sample correlations exceeding 0.3 between (a) the GCM predictor(s) so identified and (b) the station rainfall (Tippett et al. 2003).

Since CCA is susceptible to over-fitting, a simple but severe cross-validation procedure was employed, consisting of dividing the dataset into two contiguous 12-year training and validation parts. The CCA modes and NHMM model are derived from the training part of the dataset, using the ensemble-mean of the 24 GCM simulations together with the observed data. The resulting NHMM is then used to simulate rainfall occurrences for the validation part of the dataset, using the input timeseries derived by projecting the GCM ensemble-mean simulation from the validation part onto the CCA modes derived from the training part. The procedure is then repeated switching the training and validation parts of the dataset. Thus, two models are fitted and used to simulate the complementary part of the series.

In both cases the CCA yields one statistically significant mode, characterized by correlations between the canonical variate timeseries of 0.95 and 0.79, for the first and second half of the data set respectively. Figure 16 shows the structure of the CCA mode for each part of the dataset, in terms of homogeneous covariance maps formed by regressing each field with its respective (normalized) canonical variate time series. Both halves of the dataset are characterized by a correspondence between the GCM's simulated broad-scale precipitation on the southern flank of the Atlantic ITCZ and the observed station rainfall occurrence frequency over Ceará, with the relationship being stronger from 1975–1990 than for 1991–2002.

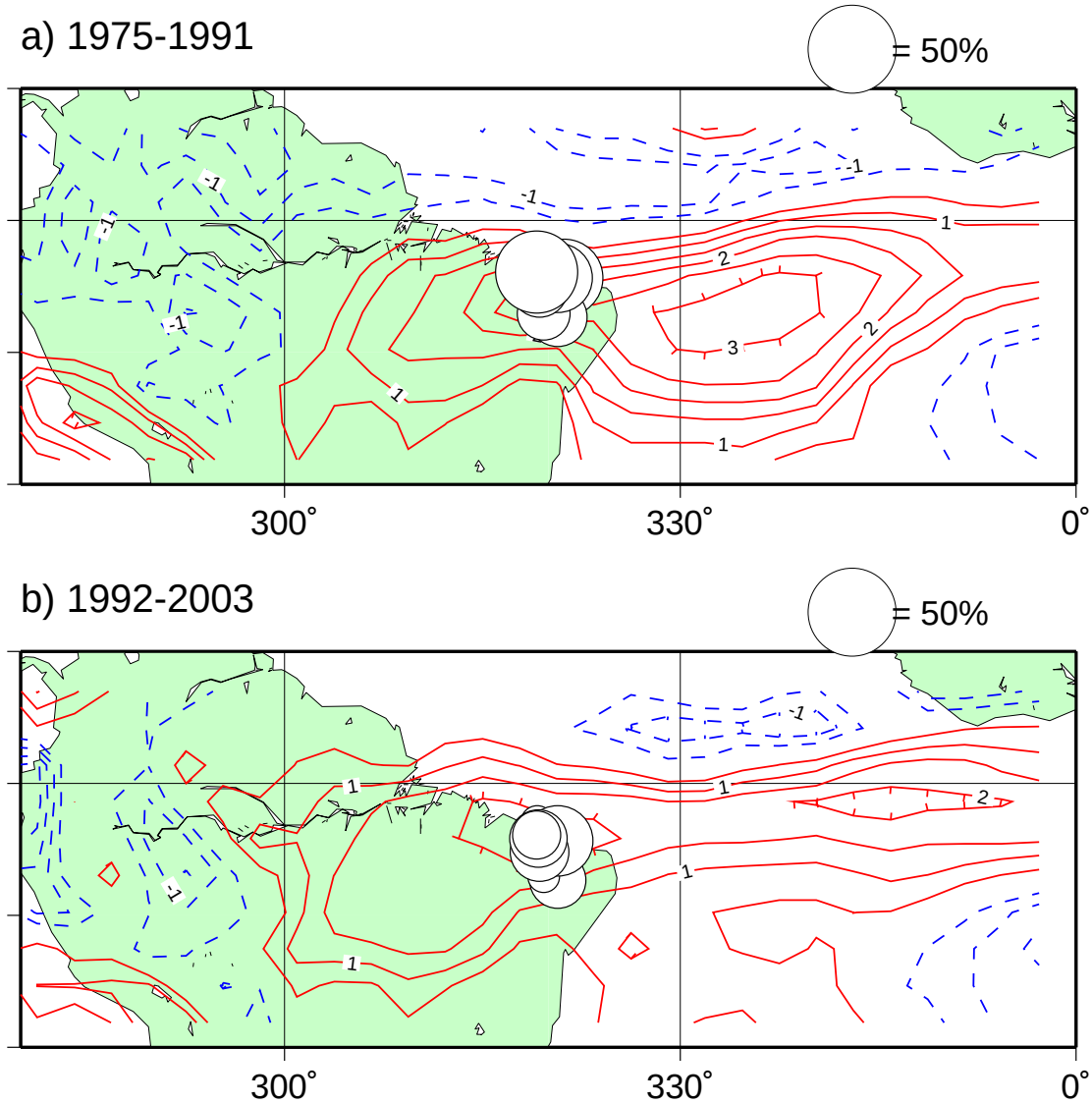

Figure 16: Canonical correlation analysis of GCM gridded seasonal-mean precipitation amount (contours) and Ceará station daily rainfall frequency (circles). The leading CCA mode is shown, computed from each half of the dataset in turn: (a) 1975–1990, (b) 1991–2002. The amplitudes correspond to a one standard deviation anomaly, with contours every 0.5. All station-rainfall anomalies are positive. Details of the CCA: The optimal PCA truncations are (a) 2 GCM PCs, 3 station PCs, and (b) 5 GCM PCs and 1 station PC, for the two halves of the data set respectively. The associated time-series correlations are (a) 0.95 and (b) 0.79.

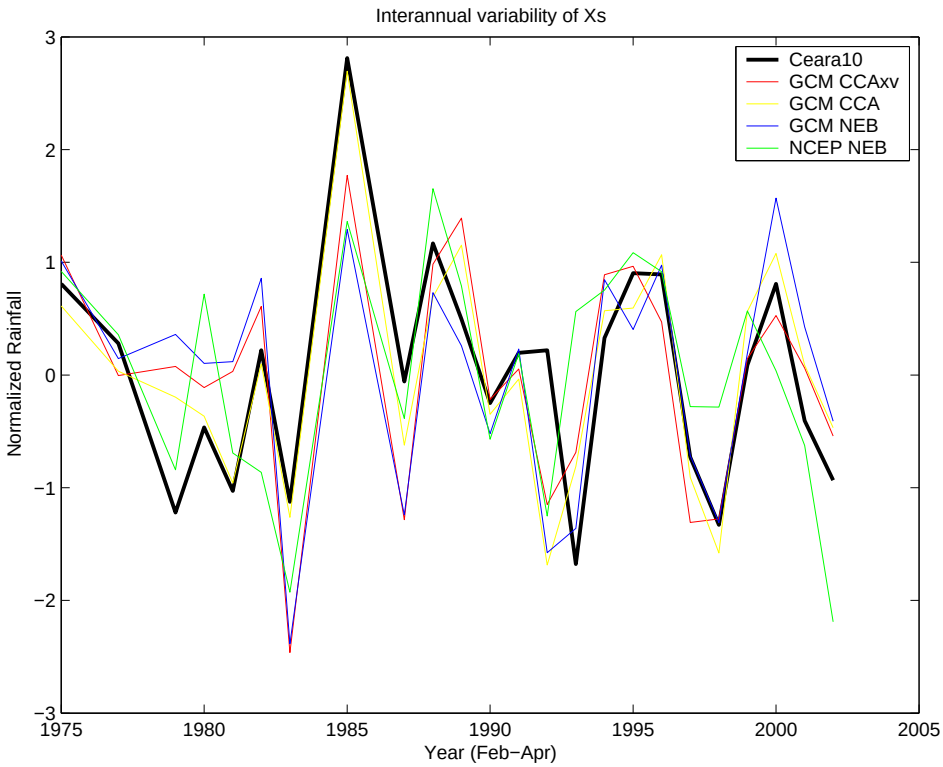

Figure 17: Interannual variability of candidate GCM predictor variables, together with the observed daily rainfall occurrence, averaged over the 10 stations (black). Key: leading cross-validated CCA mode (red, $r = 0.73$); leading CCA mode without cross-validation (yellow, $r = 0.86$); GCM NE Brazil area-average precipitation (blue, $r = 0.66$); NCEP/NCAR reanalysis area-average precipitation (green, $r = 0.62$). The reported correlation values are with the observed.

The resulting (cross-validated) GCM predictor time series is plotted in Fig. 17, together with the station-averaged observed rainfall probability; their linear correlation is 0.73. An alternative to using the CCA calibration, is to use a spatial average the GCM's seasonal-mean rainfall amount simulations over the region of NE Brazil, obviating the need for cross-validation. Figure 17 also shows these (normalized) spatial averages of NE Brazil rainfall (approx. $44^oW - 35^oW, 8^oS - 0^oN$) from both the GCM's ensemble mean ($r = 0.66$) as well as the NCEP/NCAR reanalysis ($r = 0.62$). It is encouraging that the GCM's 24-member ensemble mean performs about as well as the NCEP reanalysis. The gain achieved by calibrating the model with CCA is relatively small in this case. A predictor time series derived using CCA but without cross-validation is also plotted to illustrate the serious problem of over-fitting that results ($r = 0.86$).

5.3 NHMM simulations

The generalized EM algorithm (Appendix A) was then used to learn the parameters of the 4-state NHMM with 10 binary (rainfall) outputs and 1 real-valued (GCM) input, for each half of the dataset separately.

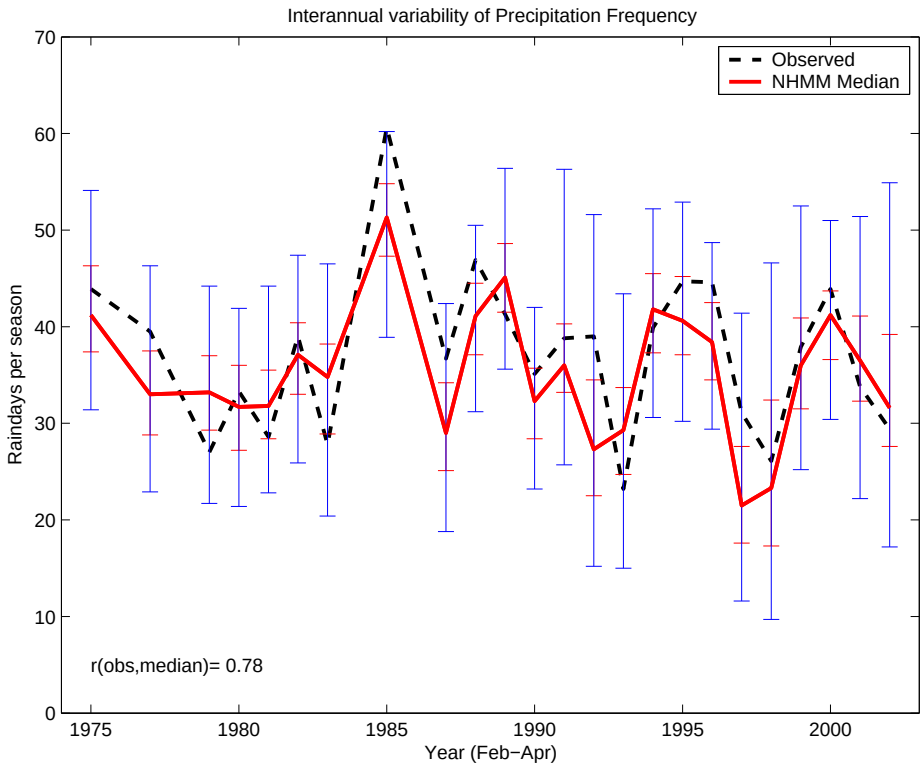

Figure 18: Interannual variability of NHMM-simulated rainfall occurrence versus the observed (black, dashed) averaged over all 10 stations. Plotted is the median of 24 NHMM simulations (red, solid). The number of raindays per season were summed across the the 10 stations, and then divided by 10. The error bars indicate the entire range of the 24 simulations, with the inter-quartile range given by the inner ticks.

Twenty-four simulations of daily rainfall occurrence were then made in each case, using the GCM ensemble mean input repeated 24 times; use of the individual GCM ensemble members was found to degrade the simulations. Figure 18 shows the median number of rain-days per season resulting from the 24 simulations, using the number of raindays averaged over all 10 stations. The linear correlation with the observed value is 0.78, which is similar to the performance of the seasonal-mean input variable in Fig. 17. Thus, the NHMM simulations recover the predictive value of the input variable in this seasonal and station average quantity. Also plotted are the quartiles and extremes of the simulated distribution. The observed curve is inside the simulated inter-quartile range about 50% of years, indicating that the simulated distribution has a consistent variance. The simulated distribution also brackets the observed one during all years, and is thus consistent with the NHMM. In other words, the NHMM is capable of generating the observed rainfall time series under strict cross-validation.

The interannual performance of the median simulated daily rainfall sequence is plotted for each station individually in Fig. 19. Stations 8 and 10 have correlations with the observed that exceed 0.7, while station 5 (on the coast) exhibits near-zero correlation. There appears to be no obvious spatial organization to the more successful interannual simulations.

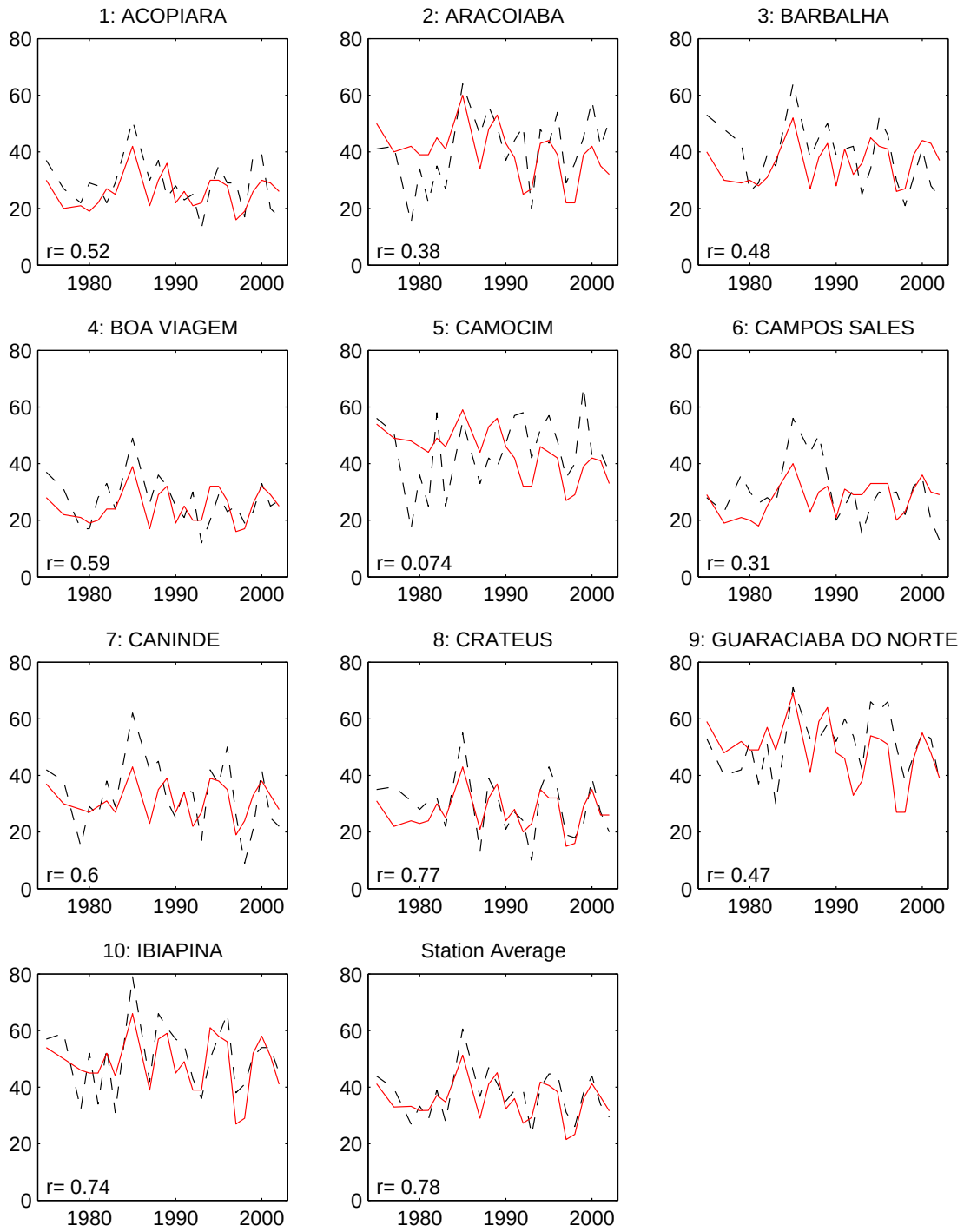

Figure 19: Interannual variability of NHMM-simulated rainfall occurrence frequency at each station. The median of the 24 simulations is plotted for each year (red, solid) together with the observed (black, dashed). The number of raindays per season is plotted on the ordinate.

The distribution of dry-spells is a parameter that is of particular importance to agriculture. Figure 20 shows the interannual variability of NHMM-simulated dry-spell frequency at each station, in terms of the median of the simulated distribution for each year. Here we define dry spells to be runs of dry days of at least 10 days, with no more than one intervening wet day. Again, rainfall amount is not considered. The GCM-NHMM is able to simulate interannual variability in dry-spell frequency fairly well at several stations, with anomaly correlations (between observed and median simulated) often approaching those for rainfall frequency.

6 Summary and Conclusions

We have used a hidden Markov model (HMM) to analyze daily rainfall occurrence at ten gauge-stations over the state of Ceará in NE Brazil during the rainy season (FMA) 1975–2002. A four-state model is chosen from inspection of the cross-validated log-likelihood of the rainfall data given the model and the Bayes Information Criterion (Table 1), as well as from subjective considerations. Unlike the BIC, the log-likelihood does not reach a peak at $K = 4$ using the leave-six-year-out cross-validation. It may be preferable to omit more years, but the dataset of only 24 complete FMA seasons is a limiting factor.

The HMM is used to estimate the hidden state sequence underlying the observed data, from which seasonal and interannual variability is analyzed. Accompanying meteorological conditions are examined through composites of NCEP/NCAR reanalysis data and interpolated NOAA OLR.

Two of the states are found to correspond to wet or dry conditions at all stations respectively, with similar relative frequencies (Fig. 7). However, the wet state (state 1) tends to be more prevalent during March, with the dry state (state 2) being more prevalent at the beginning of the FMA season (Figs. 8 and 10). Thus, on average, states 1 and 2 describe the seasonal cycle of the "monsoon" over NE Brazil, and this is brought out in the composites of anomalous OLR and low-level winds (Fig. 9). State 1 represents an anomalously southward-displaced ITCZ and an eastward expansion of the SAMS; both convergence zones merge to a greater extent than in the FMA climatology, and Ceará comes entirely under their influence. States 3 and 4 are characterized by meridional gradients of rainfall probability that have increased prevalence late in the season as the ITCZ retreats northward. However, the meteorological associations are relatively weak (cf. Fig. 9).

The state-based description of the "monsoon" over NE Brazil deserves some comment. Daily rainfall occurrence at a single station is binary, and thus it is natural to characterize it by a discrete Markov process. This is the basis of "weather generator" models of rainfall occurrence. The statistics in Table 1 demonstrate that the HMM outperforms a stateless model consisting of independent Markov chains fit to each station. The temporal evolution of the monsoon may actually be better described in terms of a discontinuous weather-state process, than by a continuous one. The state-based model provides a probabilistic description of the onset and end of the rainy season (Fig. 8), together with an average seasonal evolution (Fig. 10). Further work is needed to address this issue, using raingauge networks covering larger geographical areas that enable the spatial structures of the weather states to be better defined.

Despite the lack of any built-in rainfall persistence in the HMM, the distribution of wet and dry spell-lengths is generally reproduced surprisingly accurately. Nonetheless, a tendency to underesti-

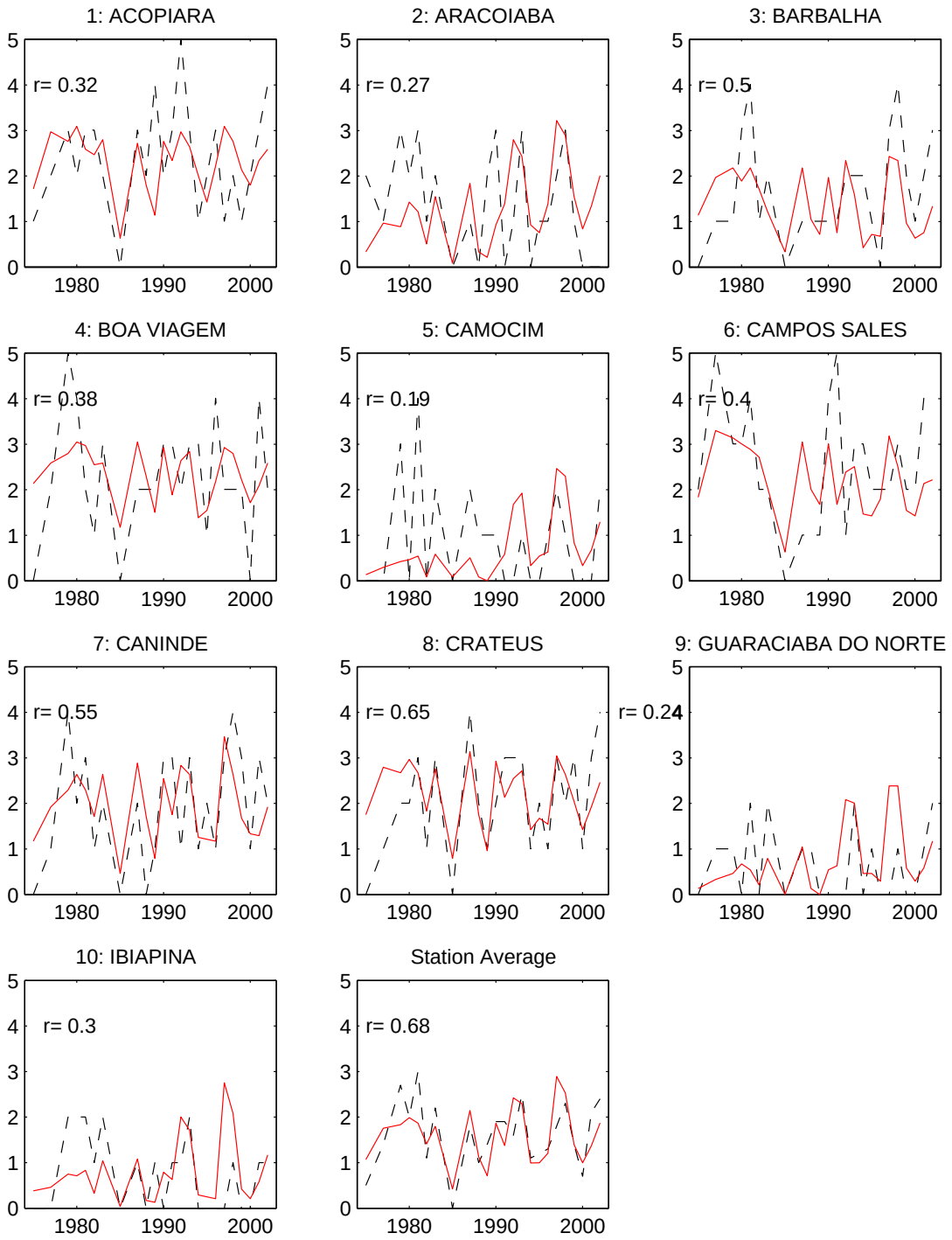

Figure 20: Interannual variability of NHMM-simulated 10-day dry-spell frequency at each station. The median of the 24 simulations is plotted for each year (red, solid) together with the observed (black, dashed). The number of dry spells per season is plotted on the ordinate.

mate the spell-lengths at certain stations is clear. An autoregressive HMM (e.g., Juang and Rabiner 1985) may provide a solution to this defect, by explicity including arrows between the outputs in Figs. 3 and 14.

The atmospheric composites of states 1 and 2 (Fig. 11) exhibit some well-known characteristics of intraseasonal (NAO) and interannual (ENSO) teleconnections. These teleconnection patterns influence the SE trade winds in the tropical Atlantic and the position of the ITCZ, with similar characteristics to the mean seasonal evolution. It is not surprising that we find indications of variability on different time scales in the rainfall states, because the atmospheric spatial structures are similar. From a regional perspective, several mechanisms can influence the probability of occurrence of the states, and thus the rainfall occurrence. Thus, the hidden states of observed rainfall occurrence appear to correspond to intrinsic weather states, which allow a natural description of rainfall variability across many different timescales.

There are large interannual variations in state-frequency, particularly of states 1 and 2, together with some decadal-scale changes (also in state 3) (Fig. 12). The SST anomalies during the years of large anomalies in state frequency (Fig. 13) bear the hallmarks of ENSO. The La Niña relationship is more statistically significant than for El Niño. This is consistent with the findings of Giannini et al. (2003) who show evidence that the relationship between NE Brazil rainfall and La Niña hs been stronger than for El Niño during recent decades, due to preconditioning by Tropical Atlantic SST anomalies. From our short 24-year daatset, we find the relationship between the latter and state-frequency to be relatively weak compared to that identified from linear correlation using longer series (e.g., Moura and Shukla 1981).

The second goal of the paper was to utilize GCM "predictions" of the large-scale circulation to make downscaled simulations of station-scale rainfall. Based on historical SST forcing, the 24-member ECHAM4.5 GCM ensemble-mean precipitation is found to have considerable cross-validated skill at reproducing interannual variations in 10-station average rainfall occurrence. We find a cross-validated linear correlation of 0.66 with the GCM's precipitation averaged over the NE Brazil region (Fig. 17). This value rises to 0.73 if a canonical correlation analysis is used to find the GCM's pattern of precipitation that is best correlated (under cross-validation) with Ceará seasonally-averaged rainfall occurrence frequency (Fig. 16).

The GCM's ensemble-mean simulation of seasonal-mean precipitation is used as a univariate input into the NHMM, from which multiple daily sequences of rainfall are then generated at the 10 stations. Validating seasonal-mean rainfall frequency simulated by the NHMM integrated across the 10 stations yields a similar skill to that reported in the previous paragraph. Thus the NHMM is shown to produce simulations of daily precipitation that are consistent with the GCM's large-scale "predictions" of interannual variations. At the individual station level, the maximum interannual correlation found between simulation and observed is 0.77, using the median of a 24-member NHMM ensemble of simulations (Fig. 19). Attempting to use the individual GCM ensemble members to make the NHMM ensemble produced an inferior result to that obtained by repeating the GCM's ensemble mean 24 times. Ten-day dry-spell frequency is also hindcast fairly well by the GCM-NHMM, with a maximum anomaly correlation skill of 0.65 at station 8 (Crateus; Fig. 20).

In terms of downscaling, we have shown that the HMM is able to quite accurately capture the characteristics of daily rainfall occurrence in terms of spell-lengths (Fig. 6) and spatial inter-

station correlations (Table 1). It is also able to convey an interannual prediction to the local scale. The method thus shows promise as a technique for generating downscaled daily rainfall-sequence scenarios for input into crop models that require such inputs. Daily rainfall amounts can be incorporated into the NHMM in a consistent manner (Charles et al. 1999, Bellone et al. 2000). A topic of future research concerns the seasonal predictability of daily rainfall intensity vs. occurrence.

Acknowledgments:

This work was supported by Department of Energy grant DE-FG02-02ER63413 and by NOAA through a block grant to the International Research Institute for Climate Prediction (IRI).

A Expectation-Maximization Algorithm for NHMMs

The precipitation occurrence data set consists of N sequences each of of length T. Let $\mathbf{R}_{nt} = \left(R_{nt}^1, \ldots, R_{nt}^M\right)$ denote a precipitation occurrence vector for day t of sequence (year) n, and let $\mathbf{X}_{nt}$ and S_{nt} be the corresponding vector of inputs and the weather state (respectively) for the same day. $\mathbf{R}_{n1:nT}$ and $\mathbf{X}_{n1:nT}$ denote sequences of precipitation and inputs for sequence (or year) n, $1 \leq n \leq N$. We assume that the observed sequences (from different years) are conditionally independent given the model. Under the NHMM model the conditional log-likelihood $l(\mathbf{\Theta})$ of the observed precipitation data, given the inputs, is defined as:

$$\begin{aligned} l(\mathbf{\Theta}) &= \ln P\left(\mathbf{R}_{11:1T}, \ldots, \mathbf{R}_{N1:NT} | \mathbf{X}_{11:1T}, \ldots, \mathbf{X}_{N1:NT}, \mathbf{\Theta}\right) \\ &= \sum_{n=1}^{N} \ln \sum_{S_{n1:nT}} P\left(S_{n1} | \mathbf{X}_{n1}, \mathbf{\Theta}\right) \prod_{t=2}^{T} P\left(S_{nt} | S_{n,t-1}, \mathbf{X}_{nt}, \mathbf{\Theta}\right) \prod_{t=1}^{T} P\left(\mathbf{R}_{nt} | S_{nt}, \mathbf{\Theta}\right). \end{aligned}$$

We seek the value of the parameter vector $\mathbf{\Theta}$ that maximizes this expression. This maximizing value cannot be obtained analytically—however, the EM algorithm provides an iterative method of climbing the $l(\mathbf{\Theta})$ surface in parameter space $\mathbf{\Theta}$. Starting with an initial set of parameters $\mathbf{\Theta}^0$, we iteratively calculate new sets of parameters improving the log-likelihood of the data at each iteration. Once a convergence criterion is reached, the last set of parameters $\hat{\mathbf{\Theta}}$ is chosen as the solution. This process of initialization followed by iterative "uphill" movement until convergence is repeated for several random initializations of $\mathbf{\Theta}^0$ and the $\hat{\mathbf{\Theta}}$ that corresponds to the largest value of $l\left(\hat{\mathbf{\Theta}}\right)$ is chosen as the maximum likelihood estimate.

In the results in this paper the probabilities defining $P\left(\mathbf{R}_t | S_t, \mathbf{\Theta}^0\right)$ were drawn from the uniform distribution on $(0, 1)$ and then normalized to satisfy the constraints $p_{sm0} + p_{sm1} = 1$. All values of the parameters used to define $P\left(S_t | S_{t-1}, \mathbf{X}_t, \mathbf{\Theta}^0\right)$ were drawn from the uniform distribution on $(0.1, 0.9)$. A change in log-likelihood of less than 0.01 from one iteration to the next of EM was used as a convergence criterion. The EM algorithm was restarted from 10 different random starting positions in parameter space and run to convergence for each starting point. The solution was the parameter vector $\mathbf{\Theta}$ that had the maximum likelihood over all 10 runs (this helps avoid poor local maxima that EM can sometimes converge to).

The updated parameters $\mathbf{\Theta}^{r+1}$ at iteration r are selected to maximize

$$\begin{aligned} Q\left(\mathbf{\Theta}^{r},\mathbf{\Theta}^{r+1}\right) &= E_{P(\mathbf{S}|\mathbf{R},\mathbf{X},\mathbf{\Theta}^{r})}\ln P\left(\mathbf{S},\mathbf{R}|\mathbf{X},\mathbf{\Theta}^{r+1}\right) \\ &= \sum_{n=1}^{N}\sum_{S_{n1:nT}} P\left(S_{n1:nT}|\mathbf{R}_{n1:nT},\mathbf{X}_{n1:nT},\mathbf{\Theta}^{r}\right)\ln P\left(\mathbf{S}_{n1:nT},\mathbf{R}_{n1:nT}|\mathbf{X}_{n1:nT},\mathbf{\Theta}^{r+1}\right). \end{aligned}$$

It can be shown that $l\left(\mathbf{\Theta}^{r+1}\right)-l\left(\mathbf{\Theta}^{r}\right)\geq Q\left(\mathbf{\Theta}^{r},\mathbf{\Theta}^{r+1}\right)-Q\left(\mathbf{\Theta}^{r},\mathbf{\Theta}^{r}\right)$, so by maximizing $Q\left(\mathbf{\Theta}^{r},\mathbf{\Theta}^{r+1}\right)$, we guarantee an improvement in log-likelihood.

$Q\left(\mathbf{\Theta}^{r},\mathbf{\Theta}^{r+1}\right)$ is maximized in two steps. In the first, the E-step, we calculate $P\left(S_{n1:nT}|\mathbf{R}_{n1:nT},\mathbf{X}_{n1:nT},\mathbf{\Theta}^{r}\right)$. In the second, the M-step, we maximize $Q\left(\mathbf{\Theta}^{r},\mathbf{\Theta}^{r+1}\right)$ with respect to the parameters in $\mathbf{\Theta}^{r+1}$. While it is infeasible to calculate and store probabilities of $N*K^{T}$ possible sequences of hidden states $P\left(S_{n1:nT}|\mathbf{R}_{n1:nT},\mathbf{X}_{n1:nT},\mathbf{\Theta}^{r}\right)$ as suggested in the E-step, it turns out we need only a manageable set of $N*T*K$ probabilities $A_{nt}\left(i\right)=P\left(S_{nt}=i|\mathbf{R}_{n1:nT},\mathbf{X}_{n1:nT},\mathbf{\Theta}^{r}\right)$ and $N*(T-1)*K^{2}$ probabilities $B_{nt}\left(i,j\right)=P\left(S_{nt}=i,S_{n,t-1}=j|\mathbf{R}_{n1:nT},\mathbf{X}_{n1:nT},\mathbf{\Theta}^{r}\right)$ is sufficient to perform optimization in the M-step:

$$\begin{aligned} Q\left(\mathbf{\Theta}^{r},\mathbf{\Theta}^{r+1}\right) &= \sum_{n=1}^{N}\sum_{S_{n1:nT}} P\left(S_{n1:nT}|\mathbf{R}_{n1:nT},\mathbf{X}_{n1:nT},\mathbf{\Theta}^{r}\right)\ln P\left(S_{n1:nT},\mathbf{R}_{n1:nT}|\mathbf{X}_{n1:nT},\mathbf{\Theta}^{r+1}\right) \\ &= \sum_{n=1}^{N}\sum_{S_{n1:nT}} P\left(S_{n1:nT}|\mathbf{R}_{n1:nT},\mathbf{X}_{n1:nT},\mathbf{\Theta}^{r}\right)\left(\sum_{t=1}^{T}\ln P\left(\mathbf{R}_{nt}|S_{nt},\mathbf{\Theta}^{r+1}\right)\right. \qquad (8) \\ &\quad \left.+\ln P\left(S_{n1}|\mathbf{X}_{n1},\mathbf{\Theta}^{r+1}\right)+\sum_{t=2}^{T}\ln P\left(S_{nt}|S_{n,t-1},\mathbf{X}_{nt},\mathbf{\Theta}^{r+1}\right)\right) \\ &= \sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{i=1}^{K} A_{nt}\left(i\right)\ln P\left(\mathbf{R}_{nt}|S_{nt}=i,\mathbf{\Theta}^{r+1}\right) \qquad (9) \\ &\quad +\sum_{n=1}^{N}\sum_{i=1}^{K} A_{n1}\left(i\right)\ln P\left(S_{n1}=i|\mathbf{X}_{n1},\mathbf{\Theta}^{r+1}\right) \qquad (10) \\ &\quad +\sum_{n=1}^{N}\sum_{t=2}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} B_{nt}\left(i,j\right)\ln P\left(S_{nt}=i|S_{n,t-1}=j,\mathbf{X}_{nt},\mathbf{\Theta}^{r+1}\right). \qquad (11) \end{aligned}$$

The quantities A_{nt} and B_{nt} can be calculated using the recursive Forward-Backward procedure (Rabiner 1989). For each value of each hidden state, we recursively calculate a summary of information preceding the state (α_{nt}) and following the state (β_{nt}) as follows:

$$\begin{aligned} \alpha_{nt}\left(i\right) &= P\left(S_{nt}=i,\mathbf{R}_{n1:nt}|\mathbf{X}_{n1:nt},\mathbf{\Theta}^{r}\right) \text{ and} \\ \beta_{nt}\left(i\right) &= P\left(\mathbf{R}_{n,t+1:nT}|S_{nt}=i,\mathbf{X}_{n,t+1:nT},\mathbf{\Theta}^{r}\right). \end{aligned}$$

Then

$$\begin{aligned}
\alpha_{n1}(i) &= P(S_{n1}=i|\mathbf{X}_{n1},\mathbf{\Theta}^r)\,P(\mathbf{R}_{n1}|S_{n1}=i,\mathbf{\Theta}^r);\\
\alpha_{n,t+1}(j) &= P(\mathbf{R}_{n,t+1}|S_{n,t+1}=j,\mathbf{\Theta})\sum_{i=1}^{K}P(S_{n,t+1}=j|S_{nt}=i,\mathbf{X}_{n,t+1},\mathbf{\Theta}^r)\,\alpha_{nt}(i);\\
\beta_{nT}(i) &= 1;\\
\beta_{nt}(i) &= \sum_{j=1}^{K}P(S_{n,t+1}=j|S_{nt}=i,\mathbf{X}_{n,t+1},\mathbf{\Theta}^r)\,P(\mathbf{R}_{n,t+1}|S_{n,t+1}=j,\mathbf{\Theta}^r)\,\beta_{n,t+1}(j).
\end{aligned}$$

Once the values of α and β are obtained, the values of A and B can be computed:

$$\begin{aligned}
A_{nt}(i) &= \frac{\alpha_{nt}(i)\,\beta_{nt}(i)}{\sum_{k=1}^{K}\alpha_{nT}(k)};\\
B_{nt}(i,j) &= \frac{P(\mathbf{R}_{nt}=\mathbf{r}_{nt}|S_{nt}=i,\mathbf{\Theta}^r)\,P(S_{nt}=i|S_{n,t-1}=j,\mathbf{X}_{nt},\mathbf{\Theta}^r)\,\alpha_{n,t-1}(j)\,\beta_{nt}(i)}{\sum_{k=1}^{K}\alpha_{nT}(k)}.
\end{aligned}$$

For the M-step, the most direct way to maximize $Q\left(\mathbf{\Theta}^r,\mathbf{\Theta}^{r+1}\right)$ is to take partial derivatives of Q (with added Lagrangians to adjust for constraints) with respect to parameters in $\mathbf{\Theta}^{r+1}$ and to make all of the partial derivatives zero. This approach yields a closed form solution for the parameters of $P(\mathbf{R}_t|S_t)$:

$$\begin{aligned}
\hat{p}_{im0} &= \frac{\sum_{r_{nt}^m=0}A_{nt}(i)}{\sum_{n=1}^{N}\sum_{t=1}^{T}A_{nt}(i)} = \frac{\sum_{r_{nt}^m=0}\frac{1}{L_n}\alpha_{nt}(i)\,\beta_{nt}(i)}{\sum_{n=1}^{N}\frac{1}{L_n}\sum_{t=1}^{T}\alpha_{nt}(i)\,\beta_{nt}(i)};\\
\hat{p}_{im1} &= \frac{\sum_{r_{nt}^m=1}A_{nt}(i)}{\sum_{n=1}^{N}\sum_{t=1}^{T}A_{nt}(i)} = \frac{\sum_{r_{nt}^m=1}\frac{1}{L_n}\alpha_{nt}(i)\,\beta_{nt}(i)}{\sum_{n=1}^{N}\frac{1}{L_n}\sum_{t=1}^{T}\alpha_{nt}(i)\,\beta_{nt}(i)},
\end{aligned}$$

where $L_n=\sum_{k=1}^{K}\alpha_{nT}(k)$.

Unfortunately, parameters of the transition $P(S_t|S_{t-1},\mathbf{X}_t)$ have non-linear partial derivatives, and the system equations resulting from equating their partial derivatives to zero cannot be solved analytically. Hughes et al. (1999) resort to a black-box general numerical optimization package to solve for the parameters of their model numerically. This method can be quite slow in practice. We implemented an alternative conjugate gradient algorithm tailored for the NHMM model for the results reported in this paper. This conjugate gradient optimization is run within each iteration (during the M-step) of the EM algorithm.

The Conjugate Gradient Algorithm

The conjugate gradient method iteratively identifies directions in the space of the parameters and maximizes the objective function along each of the directions. The directions are chosen such that each successive direction is orthogonal to the gradient of the previous estimate of the parameters and conjugate to all previous directions, thus, reducing the overlap in the optimization space from iteration to iteration. In order, to apply the conjugate gradient algorithm, we need to be able to find a gradient vector for a set of parameters and to optimize the objective function along an arbitrary vector, i.e., to solve a linear search problem.

From the equation defining Q (Equation 8), note that Q can be optimized with respect to the parameters of $P(\mathbf{R}_t|S_t)$ (line 9) independently of the parameters of $P(S_t|S_{t-1}, \mathbf{X}_t)$ (lines 10 and 11). Let

$$\mathbf{\Theta}_S = (\lambda_1, \ldots, \lambda_K, \sigma_{11}, \ldots, \sigma_{KK}, \rho_1, \ldots, \rho_K)$$

be the vector of parameters for the transition probabilities $P(S_t|S_{t-1}, \mathbf{X}_t)$, and let $Q_S\left(\mathbf{\Theta}_S^r, \mathbf{\Theta}_S^{r+1}\right)$ be the terms of $Q\left(\mathbf{\Theta}^r, \mathbf{\Theta}^{r+1}\right)$ involving the parameters of $\mathbf{\Theta}_S$. The conjugate gradient method is needed to optimize Q_S.

A new set of parameters $\mathbf{\Theta}_S^{r+1}$ will be obtained from the old $\mathbf{\Theta}_S^r$ in an iterative manner. Starting with $\mathbf{\Theta}_0 = \mathbf{\Theta}_S^r$, each successive $\mathbf{\Theta}_{l+1}$ is a set of parameters derived from the previous set of parameters $\mathbf{\Theta}_l$ by optimizing Q_S in the direction $\mathbf{\Phi}_l$, i.e.

$$\mathbf{\Theta}_{l+1} = \mathbf{\Theta}_l + \nu_l \mathbf{\Phi}_l \text{ where } \nu_l = \arg\max_{\nu} Q_S\left(\mathbf{\Theta}_S^r, \mathbf{\Theta}_l + \nu \mathbf{\Phi}_l\right). \tag{12}$$

Note that for all $l \geq 0$, $Q_S\left(\mathbf{\Theta}_S^r, \mathbf{\Theta}_{l+1}\right) \geq Q_S\left(\mathbf{\Theta}_S^r, \mathbf{\Theta}_l\right) \geq Q_S\left(\mathbf{\Theta}_S^r, \mathbf{\Theta}_S^r\right)$; thus by choosing $\mathbf{\Theta}_S^{r+1} = \mathbf{\Theta}_l$ for any positive l would yield an improvement in log-likelihood. Usually, final l is chosen such that the objective function does not significantly change between $\mathbf{\Theta}_{l+1}$ and $\mathbf{\Theta}_l$.

We need to specify how to choose directions $\mathbf{\Phi}_l$. The standard gradient descent algorithm uses gradients $\mathbf{\Phi}_l = \nabla(\mathbf{\Theta}_l)$; it is, however, usually slow to converge since each of the new directions could have significant overlap with previously chosen direction vectors. Conjugate gradient method reduces the overlap between the optimization directions and can, in turn, speed up convergence to a solution. We used the Polak-Ribiere variation of the conjugate gradient method (Press et al. 1992, Chapter 10.6):

$$\begin{aligned}
\mathbf{\Phi}_0 &= -\nabla(\mathbf{\Theta}_0); \\
\mathbf{\Phi}_{l+1} &= \nabla(\mathbf{\Theta}_l) - \gamma_l \mathbf{\Phi}_l; \\
\gamma_l &= \frac{\left(\nabla(\mathbf{\Theta}_{l+1}) - \nabla(\mathbf{\Theta}_l)\right) \cdot \nabla(\mathbf{\Theta}_{l+1})}{\nabla(\mathbf{\Theta}_l) \cdot \nabla(\mathbf{\Theta}_l)}.
\end{aligned}$$

All that remains is to perform the line search to find ν_l in Equation 12. This can be accomplished by any line optimization algorithm. We solve it via Newton-Raphson by finding the zero of the derivative of $Q_S\left(\mathbf{\Theta}_S^r, \mathbf{\Theta}_l + \nu_l \mathbf{\Phi}_l\right)$ with respect to ν_l. We also need to calculate first and second derivatives

of $Q_S\left(\boldsymbol{\Theta}_S^r, \boldsymbol{\Theta}_l + \nu_l \boldsymbol{\Phi}_l\right)$ with respect to ν_l. Assuming $\boldsymbol{\Phi}_l = (\lambda_{1\phi}, \ldots, \lambda_{K\phi}, \sigma_{11\phi}, \ldots, \sigma_{KK\phi}, \rho_{1\phi}, \ldots, \rho_{K\phi})$,

$$\begin{aligned}
\frac{\mathrm{d}Q_S\left(\boldsymbol{\Theta}_S^r, \boldsymbol{\Theta}_l + \nu_l \boldsymbol{\Phi}_l\right)}{\mathrm{d}\nu_l} &= \sum_{n=1}^{N}\sum_{i=1}^{K} A_{n1}(i)\left(\lambda_{i\phi} + \boldsymbol{\rho}'_{i\phi}\mathbf{x}_{n1}\right)\left(1 - P\left(S_{n1} = i|\mathbf{X}_{n1}, \boldsymbol{\Theta}_l + \nu_l\boldsymbol{\Phi}_l\right)\right) \\
&\quad + \sum_{n=1}^{N}\sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{i=1}^{K} B_{nt}(i,j)\left(\sigma_{ji\phi} + \boldsymbol{\rho}'_{i\phi}\mathbf{x}_{n1}\right) \times \\
&\quad \times \left(1 - P\left(S_{nt} = i|S_{n,t-1} = j, \mathbf{X}_{nt}, \boldsymbol{\Theta}_l + \nu_l\boldsymbol{\Phi}\right)\right); \\
\frac{\mathrm{d}^2Q_S\left(\boldsymbol{\Theta}_S^r, \boldsymbol{\Theta}_l + \nu_l \boldsymbol{\Phi}_l\right)}{\mathrm{d}\nu_l^2} &= -\sum_{n=1}^{N}\sum_{i=1}^{K} A_{n1}(i)\left(\lambda_{i\phi} + \boldsymbol{\rho}'_{i\phi}\mathbf{x}_{n1}\right)^2 \times \\
&\quad \times P\left(S_{n1} = i|\mathbf{X}_{n1}, \boldsymbol{\Theta}_l + \nu_l\boldsymbol{\Phi}_l\right)\left(1 - P\left(S_{n1} = i|\mathbf{X}_{n1}, \boldsymbol{\Theta}_l + \nu_l\boldsymbol{\Phi}_l\right)\right) \\
&\quad - \sum_{n=1}^{N}\sum_{t=2}^{T}\sum_{j=1}^{K}\sum_{i=1}^{K} B_{nt}(i,j)\left(\sigma_{ji\phi} + \boldsymbol{\rho}'_{i\phi}\mathbf{x}_{n1}\right)^2 \times \\
&\quad \times P\left(S_{nt} = i|S_{n,t-1} = j, \mathbf{X}_{nt}, \boldsymbol{\Theta}_l + \nu_l\boldsymbol{\Phi}_l\right) \times \\
&\quad \times \left(1 - P\left(S_{nt} = i|S_{n,t-1} = j, \mathbf{X}_{nt}, \boldsymbol{\Theta}_l + \nu_l\boldsymbol{\Phi}_l\right)\right).
\end{aligned}$$

This completes the information required to implement conjugate gradient for the M-step of an EM algorithm for an NHMM model.

B BIC Scores

The BIC score for an HMM or NHMM model with K states is defined as

$$\mathrm{BIC}_K = 2L\left(\boldsymbol{\Theta}_{\mathbf{K}}^*\right) - p\log T$$

where $\boldsymbol{\Theta}_{\mathbf{K}}^*$ is the estimated maximum likelihood parameter vector as found by EM on the training data for a model with K states, $L\left(\boldsymbol{\Theta}_{\mathbf{K}}^*\right)$ is the likelihood of the model evaluated at $\boldsymbol{\Theta}_{\mathbf{K}}^*$ as in Equation 7, p is the number of parameters in the K-state model (linear in K), and T is the total number of days of observed data used to train the model. The second term in the BIC expression $-p\log T$ "penalizes" more complex models. BIC can be viewed as a practical approximation to the more ideal (but intractable to compute) Bayes factor for model selection (e.g., see Kass and Raftery 1995). Although not fully justified theoretically for model selection in the context of HMMs and NHMMs (e.g., see Titterington (1990) and Hughes et al. (1999) for further comments), the BIC score can nonetheless provide a useful indication of which models are supported by the data (e.g., Hughes and Guttorp 1994).

To obtain normalized BIC scores (as in Table 1) that are on roughly the same scale as the normalized log-likelihoods, we replace BIC_K above with $-\mathrm{BIC}_K/2N$, where N is the total number of binary predictions made (here $N = 24 \times 90 \times 10$).

References

F. Bauer. Extended range weather forecasting. In *Compendium of Meteorology*, pages 814–833. American Meteorological Society, Boston, 1951.

E. Bellone, J. P. Hughes, and P. Guttorp. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Climate Research*, 15(1):1–12, May 15 2000.

S. P. Charles, B. C. Bates, and J. P. Hughes. A spatiotemporal model for downscaling precipitation occurrence and amounts. *Journal of Geophysical Research–Atmospheres*, 104(D24):31657–31669, December 27 1999.

J. G. Charney and J. G. DeVore. Multiple flow equilibria in the atmosphere and blocking. *Journal of the Atmospheric Sciences*, 36(7):1205–1216, 1979.

J. C. H. Chiang, Y. Kushnir, and A. Giannini. Deconstructing Atlantic Intertropical Convergence Zone variability: Influence of the local cross-equatorial sea surface temperature gradient and remote forcing from the eastern equatorial Pacific. *Journal of Geophysical Research–Atmospheres*, 107(D1-D2), January 2002. Article №4004.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, 39(1):1–38, 1977.

S. B. Feldstein. The timescale, power spectra, and climate noise properties of teleconnection patterns. *Journal of Climate*, 13(24):4430–4440, 2000.

A. Giannini, R. Saravanan, and P. Chang. The preconditioning role of Tropical Atlantic variability in the prediction of Nordeste rainfall during ENSO events. *Climate Dynamics*, 2003. submitted.

L. Goddard, A. G. Barnston, and S. J. Mason. Evaluation of the IRI's "net assessment" seasonal climate forecasts: 1997–2001. *Bulletin of the American Meteorological Society*, 2003. in press.

S. Hastenrath and L. Heller. Dynamics of climatic hazards in northeast Brazil. *Quarterly Journal of the Royal Meteorological Society*, 103(435):77–92, 1977.

J. P. Hughes and P. Guttorp. Incorporating spatial dependence and atmospheric data in a model of precipitation. *Journal of Applied Meteorology*, 33(12):1503–1515, December 1994.

J. P. Hughes, P. Guttorp, and S. P. Charles. A non-homogeneous hidden Markov model for precipitation occurrence. *Journal of the Royal Statistical Society Series C Applied Statistics*, 48(1): 15–30, 1999.

J. W. Hurrell, Y. Kushnir, G. Ottersen, and M. Visbeck, editors. *The North Atlantic Oscillation : Climatic Significance and Environmental Impact*. Number 134 in Geophysical Monograph Series. American Geophysical Union, 2002.

A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ, 1988.

B. H. Juang and L. R. Rabiner. Mixture autoregresive hidden markov models for speech signals. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(6):1404–1413, 1985.

E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471, March 1996.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90 (430):773–795, June 1995.

B. Liebmann, G. N. Kiladis, J. A. Marengo, T. Ambrizzi, and J. D. Glick. Submonthly convective variability over South America and the South Atlantic convergence zone. *Journal of Climate*, 12 (7):1877–1891, July 1999.

E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963.

I. L. MacDonald and W. Zucchini. *Hidden Markov and Other Models for Discrete-valued Time Series*. Number 70 in Monographs on Statistics and Applied Probability. Chapman & Hall, 1997.

A. D. Moura and J. Shukla. On the dynamics of droughts in northeast Brazil – observations, theory and numerical experiments with a general-circulation model. *Journal of the Atmospheric Sciences*, 38(12):2653–2657, 1981.

P. Nobre and J. Shukla. Variations of sea surface temperature, wind stress, and rainfall over the tropical Atlantic and South America. *Journal of Climate*, 9(10):2464–2479, October 1996.

W. H. Press, S. A. Teukolsky, W. V. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2nd edition, 1992.

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77(2):257–286, February 1989.

E. Roeckner and Coauthors. The atmospheric geneal circulation model ECHAM4: Model description and simulation of present-day climate. *Max-Planck-Institut fur Meteorologie Rept.*, 23:90pp, 1996.

L. Sun, D. F. Moncunill, H. Li, A. D. Moura, and F. A. D. S. Filho. Climate downscaling over Nordeste Brazil using NCEP RSM97. *Journal of Climate*, 2003. in preparation.

D. W. J. Thompson, M. P. Baldwin, and J. M. Wallace. Stratospheric connection to Northern Hemisphere wintertime weather: Implications for prediction. *Journal of Climate*, 15(12):1421–1428, June 2002.

M. K. Tippett, M. Barlow, and B. Lyon. Statistical correction of central southwest asia winter precipitation simulations. *International Journal of Climatology*, 23:1421–1433, 2003.

D. M. Titterington. Some recent research in the analysis of mixture distributions. *Statistics*, 21: 619–641, 1990.

A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

D. S. Wilks and R. L. Wilby. The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, 23(3):329–357, 1999.