Markov random field model for the Indian monsoon rainfall

Adway Mitra,
Amit Apte, Rama Govindarajan, Vishal Vasan, Sreekar Vadlamani

Thanks:
Infosys Foundation;
Airbus Chair program at ICTS and CAM, TIFR;

30 June 2018
DCS-2018, ICTS-TIFR, Bangalore

Outline

Monsoon: a brief description

Markov random field (MRF) model

Results: prominent spatial patterns

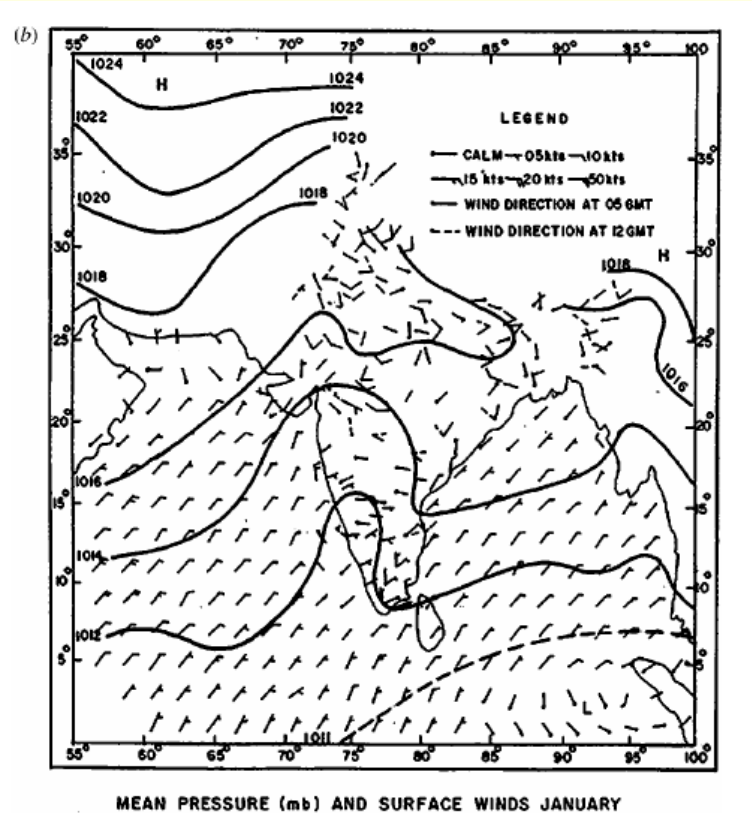
Discussion

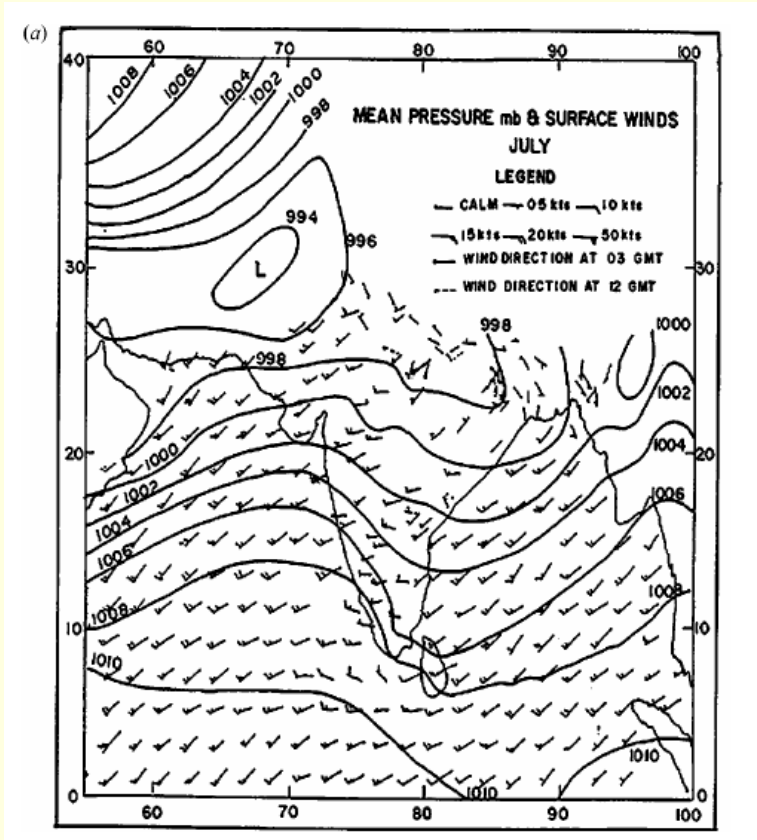
Main references for the first section:

- Gadgil, Annu. Rev. Earth Planet. Sci. 2003. 31:42967
- `http://nptel.ac.in/courses/119108006/`
- Gadgil, Nanjundiah, Srinivasan, Private communication!!
- Some plots and data:
  `http://www.tropmet.res.in/~kolli/mol/Monsoon/` and
  `http://apdrc.soest.hawaii.edu/projects/monsoon/`

Monsoon: periodic variations of wind and rain

The winds (used to) matter a lot to the sailors (who coined the term "monsoon")

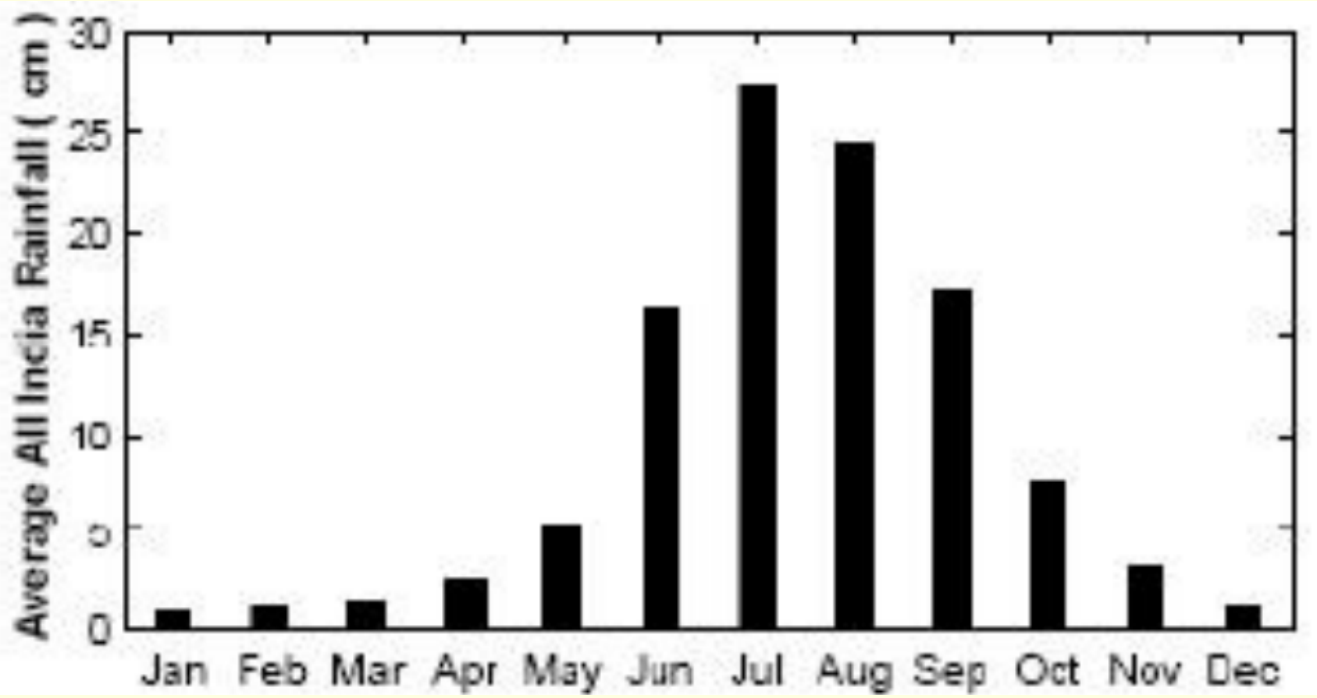




- Northeasterly winds during winter months (left)
- Southwesterly winds during summer months (right)

Monsoon: periodic variations of wind and rain

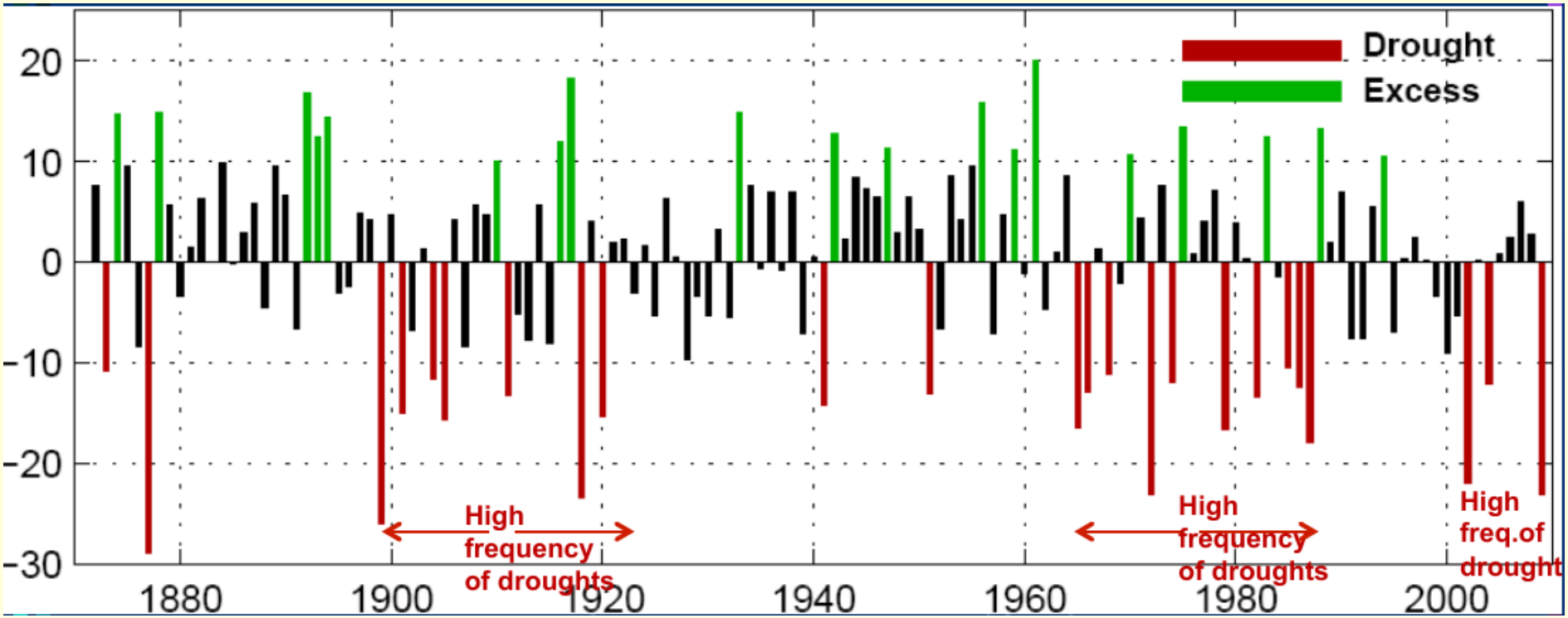
Rain is what matters most to inhabitants of the subcontinent![1]



- Hardly any rainfall during Dec-Apr
- Most of the rainfall during Jun-Sep – the monsoon season

[1] School summer vacations: april-may;
Three seasons in Mumbai: summer, rainy season, winter

Monsoon rains are quite reliable!



- Long term mean is 850 mm / four months
- Standard deviation is 10% of the mean
- ISMR[2] $<$ 90% - drought; $>$ 110% - excess; otherwise normal

[2] ISMR = Indian Summer Monsoon Rainfall

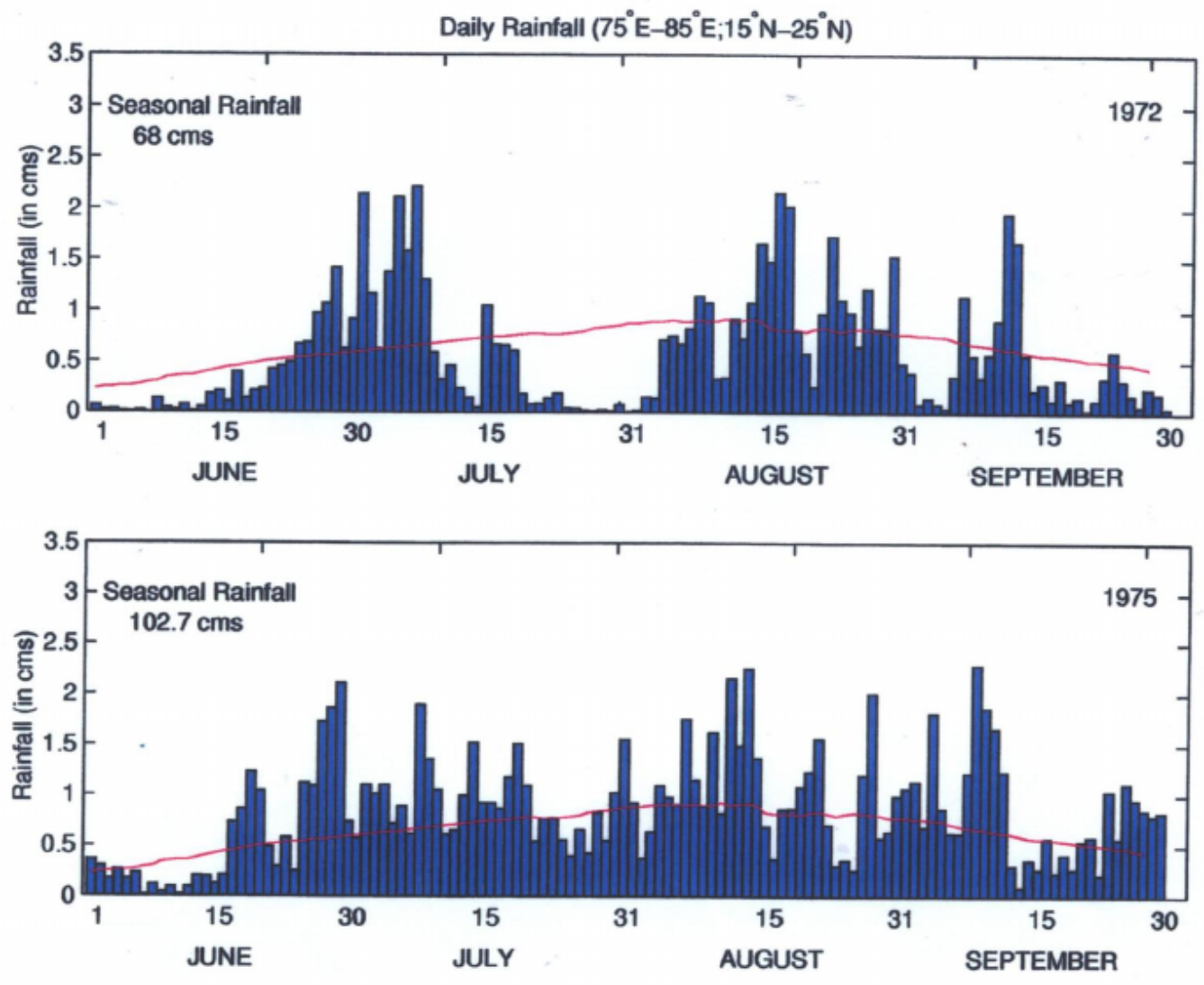
First hypothesis: land-sea temperature contrast causes monsoon rain

- Halley 1686: The primary cause of the monsoon is the differential heating between ocean and land.
- Not supported by more extensive observational data (e.g. Kothawale, Rupakumar 2002)
  - temperature contrast is higher in May (before rains) than in monsoon months
  - surface temperature anomaly is positive for droughts and negative for excess monsoon seasons
- The issue is still not completely settled. (I know of no "simple" dynamical models for this hypothesis.)

We will move on to discussing intra-seasonal variations.

There are large intraseasonal variations

top figure: drought year; bottom: excess / normal



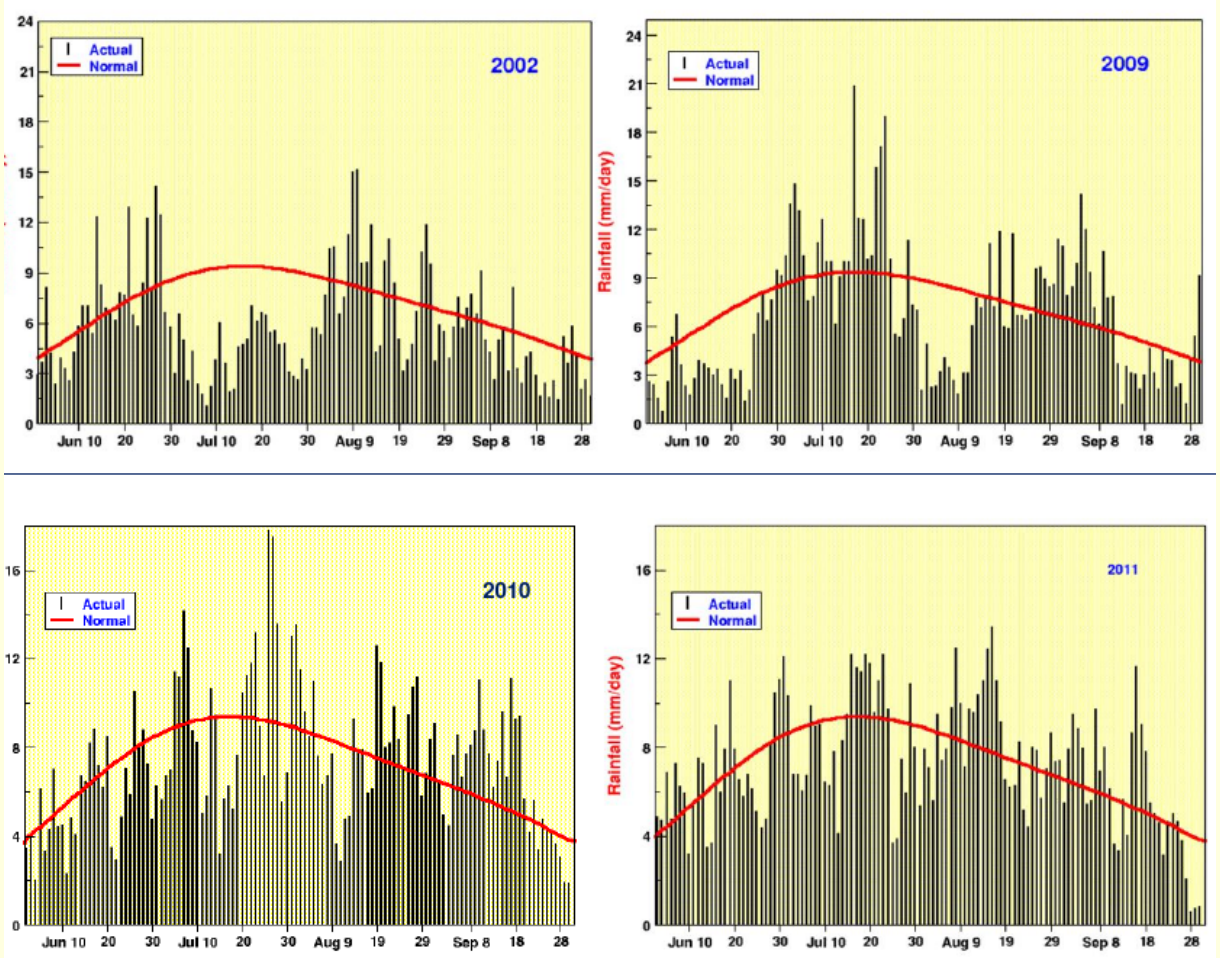
The smooth long term mean is of course because of averaging!

Most years have active and break spells during the monsoon season (no year is like the average!)

Thin line: long-term mean (first slide)

There are large intraseasonal variations

top figure: drought year; bottom: excess / normal

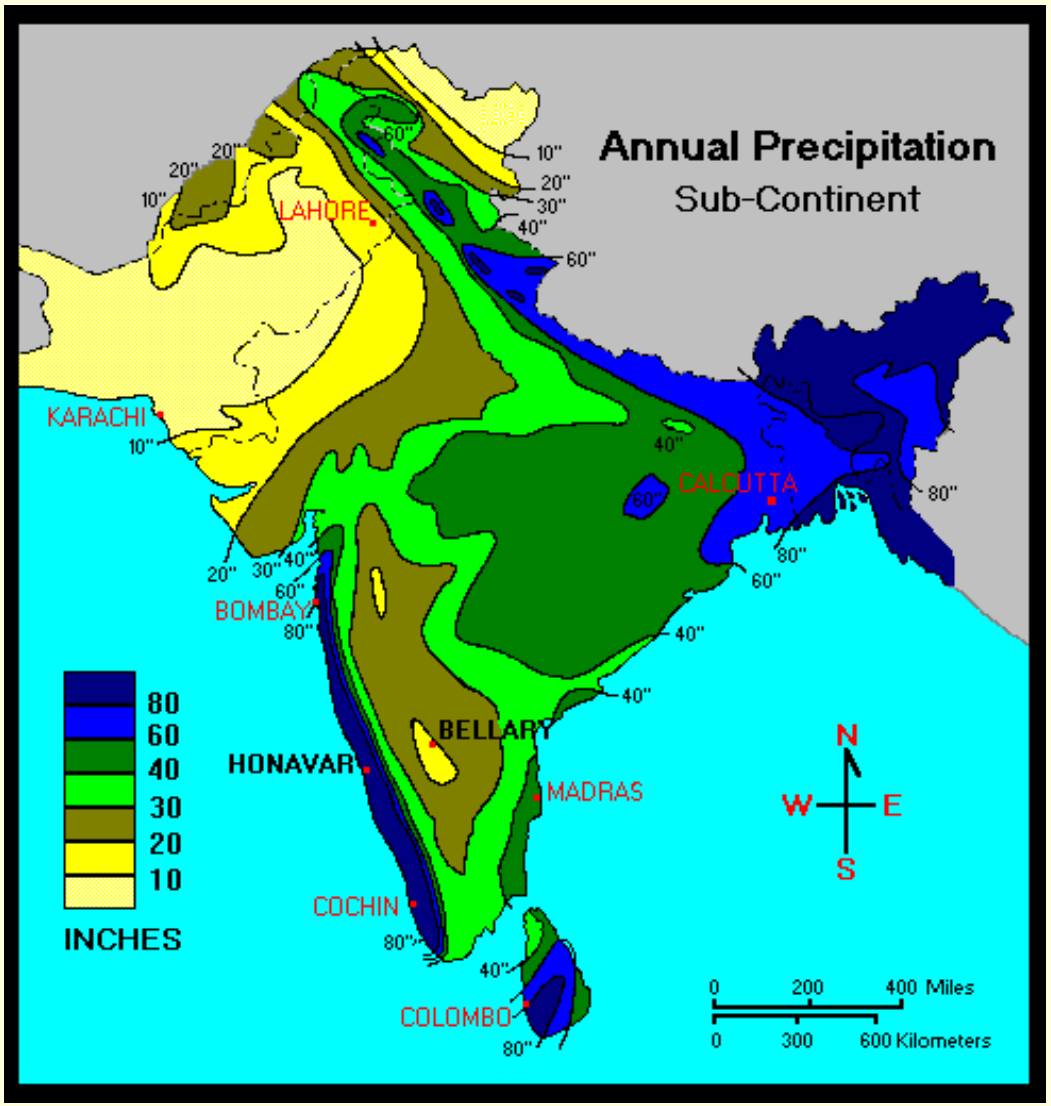


The smooth long term mean is of course because of averaging!

Most years have active and break spells during the monsoon season (no year is like the average!)

Thin line: long-term mean (first slide)

There is substantial geographic variation

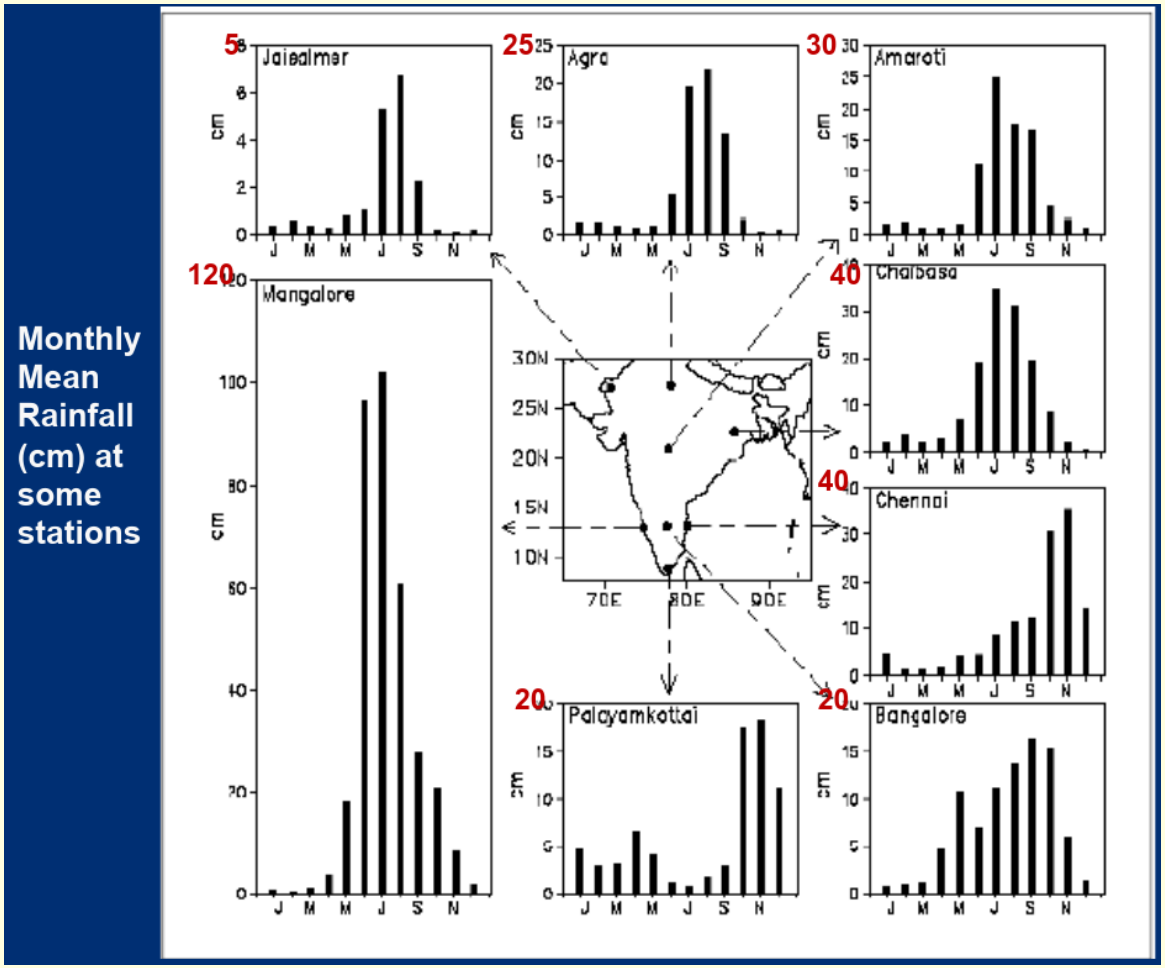


http://www.oneonta.edu/faculty/baumanpr/geosat2/Dry_Monsoon/Dry_Monsoon.htm

- Himalaya: average elevation around 5km
- Western Ghats: average elevation around 1km
- Other smaller mountain ranges

Central India is usually called the "monsoon zone."

There is substantial geographic variation



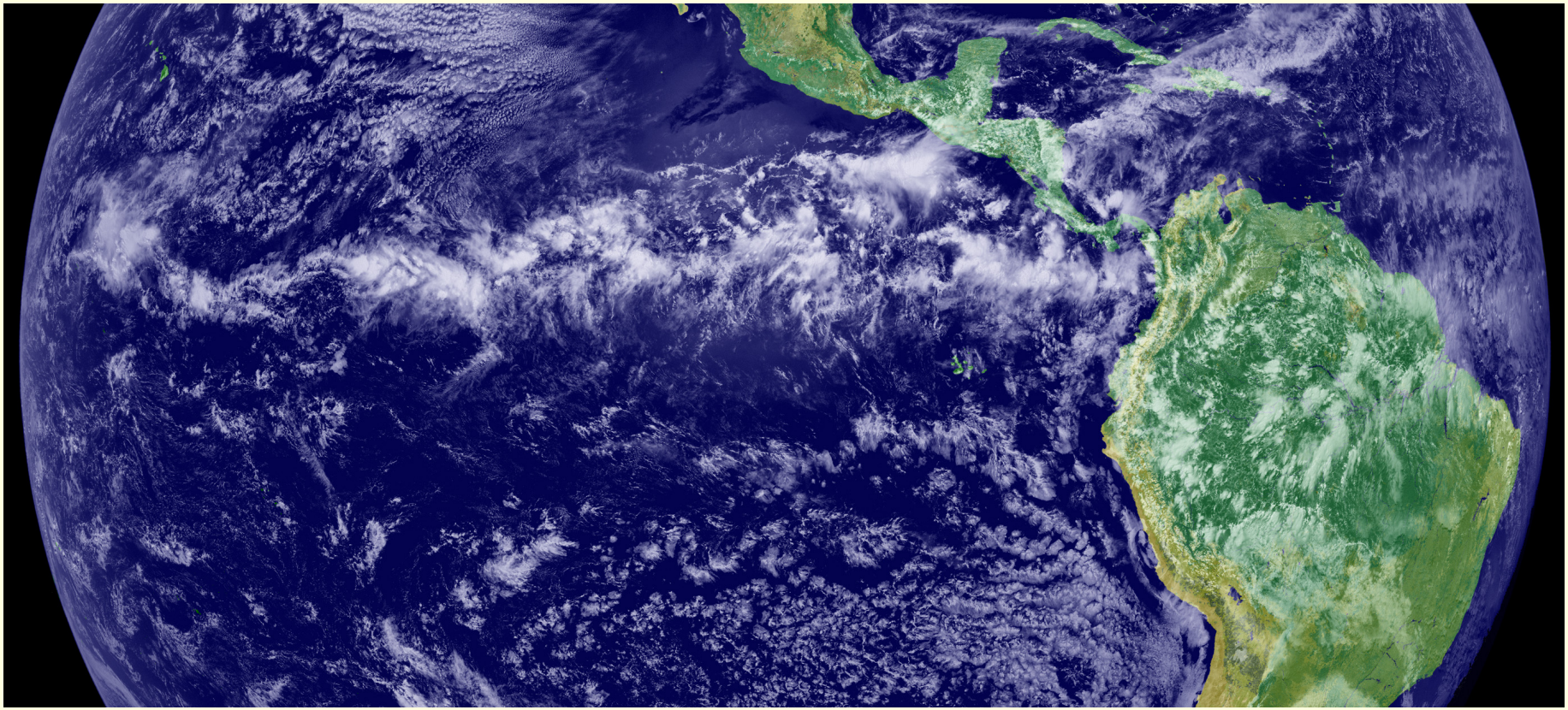
http://nptel.ac.in/courses/119108006/downloads/Lecture01.pdf

- Himalaya: average elevation around 5km
- Western Ghats: average elevation around 1km
- Other smaller mountain ranges

Central India is usually called the "monsoon zone."

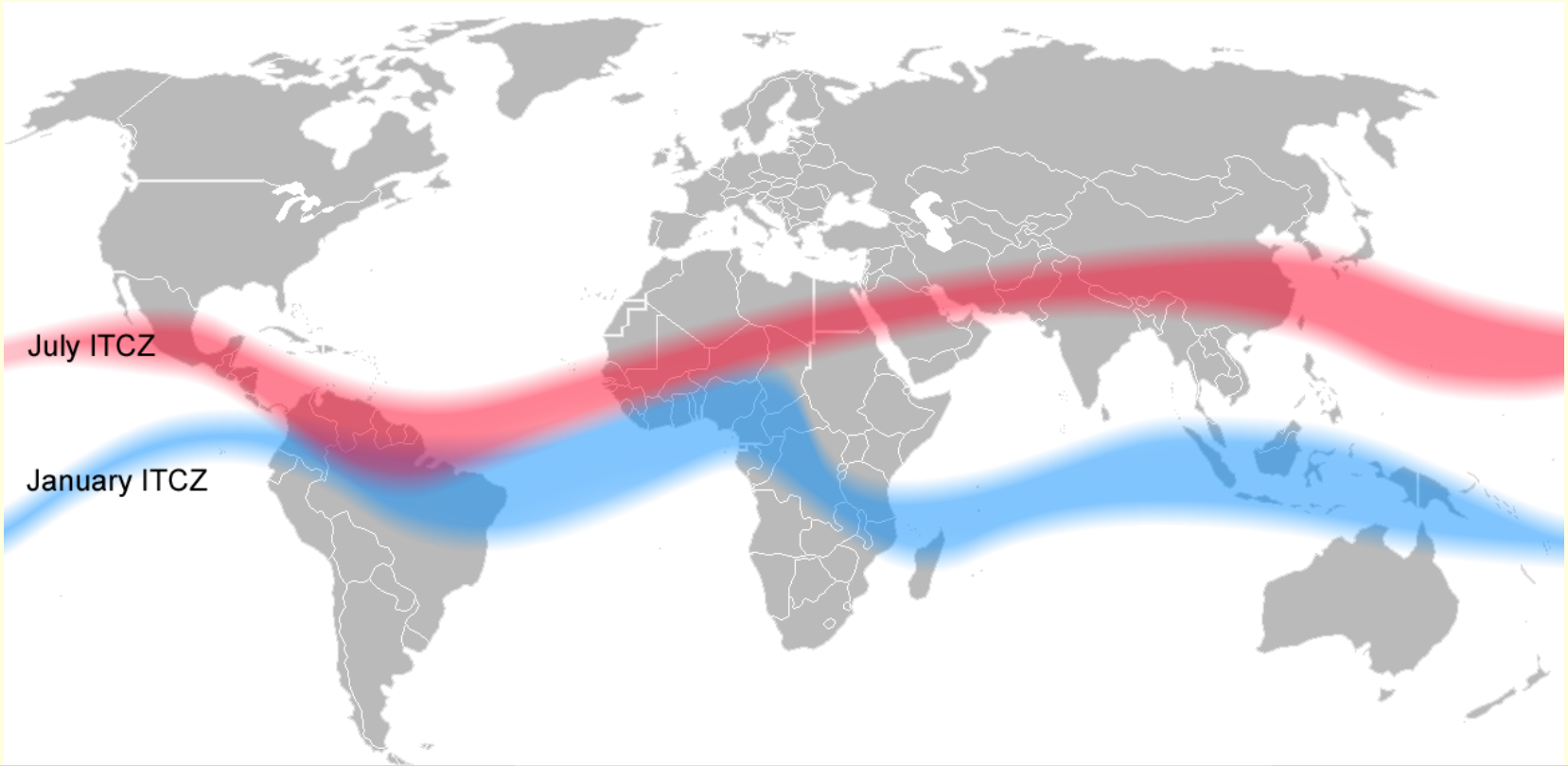
The second hypothesis: seasonal variation of ITCZ

Intertropical Convergence Zone (ITCZ): band of clouds near equatorial region across the globe

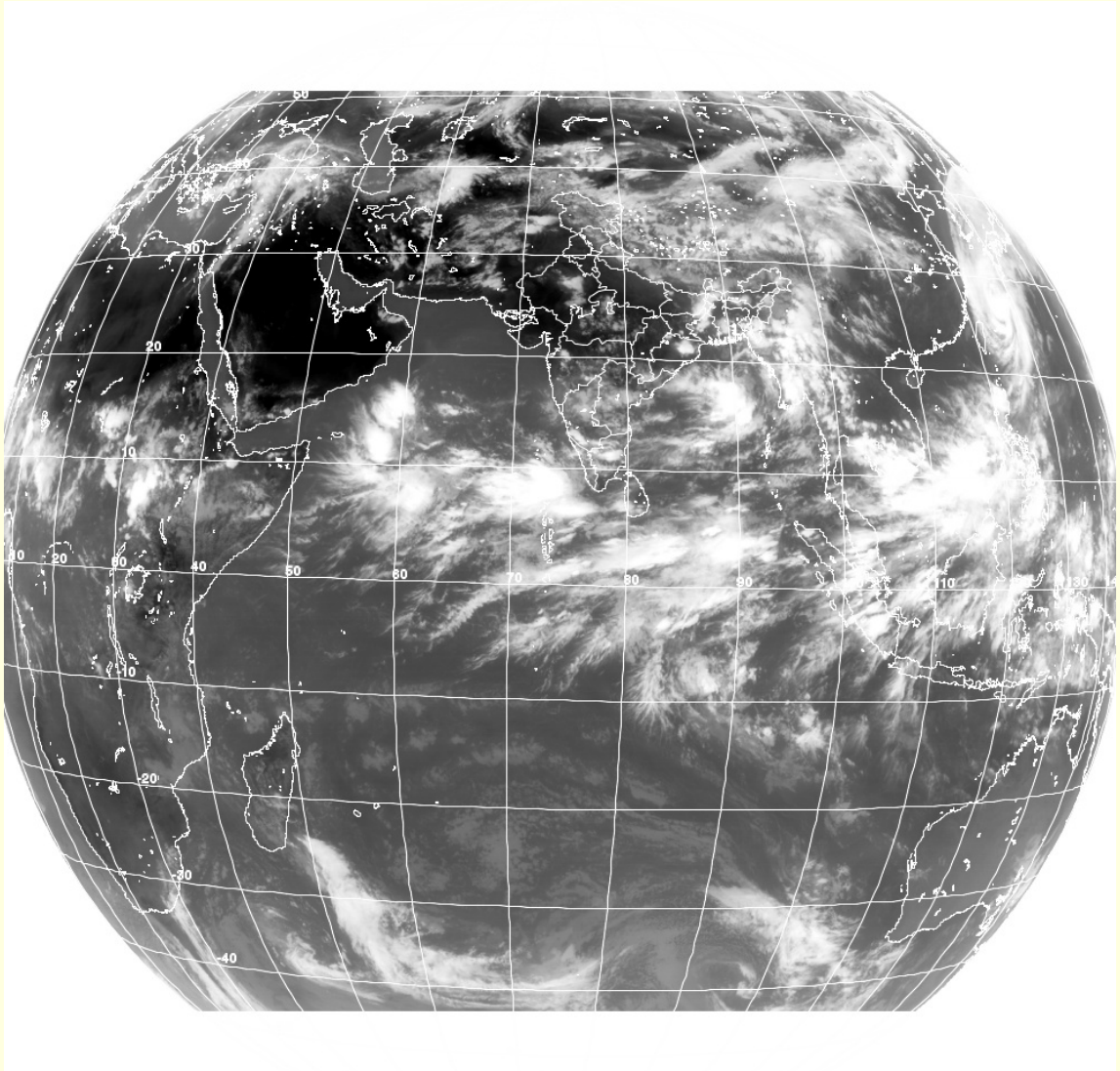


The second hypothesis: seasonal variation of ITCZ

- Charney / Riehl / Gadgil / Sikka, etc. (1970s): monsoon is attributed to the seasonal migration of the "ITCZ"
- There may be two tropical convergence zones (TCZ) over Indian subcontinent during monsoon season
- Northward propagation of the TCZ plays important role in the intra-seasonal variation

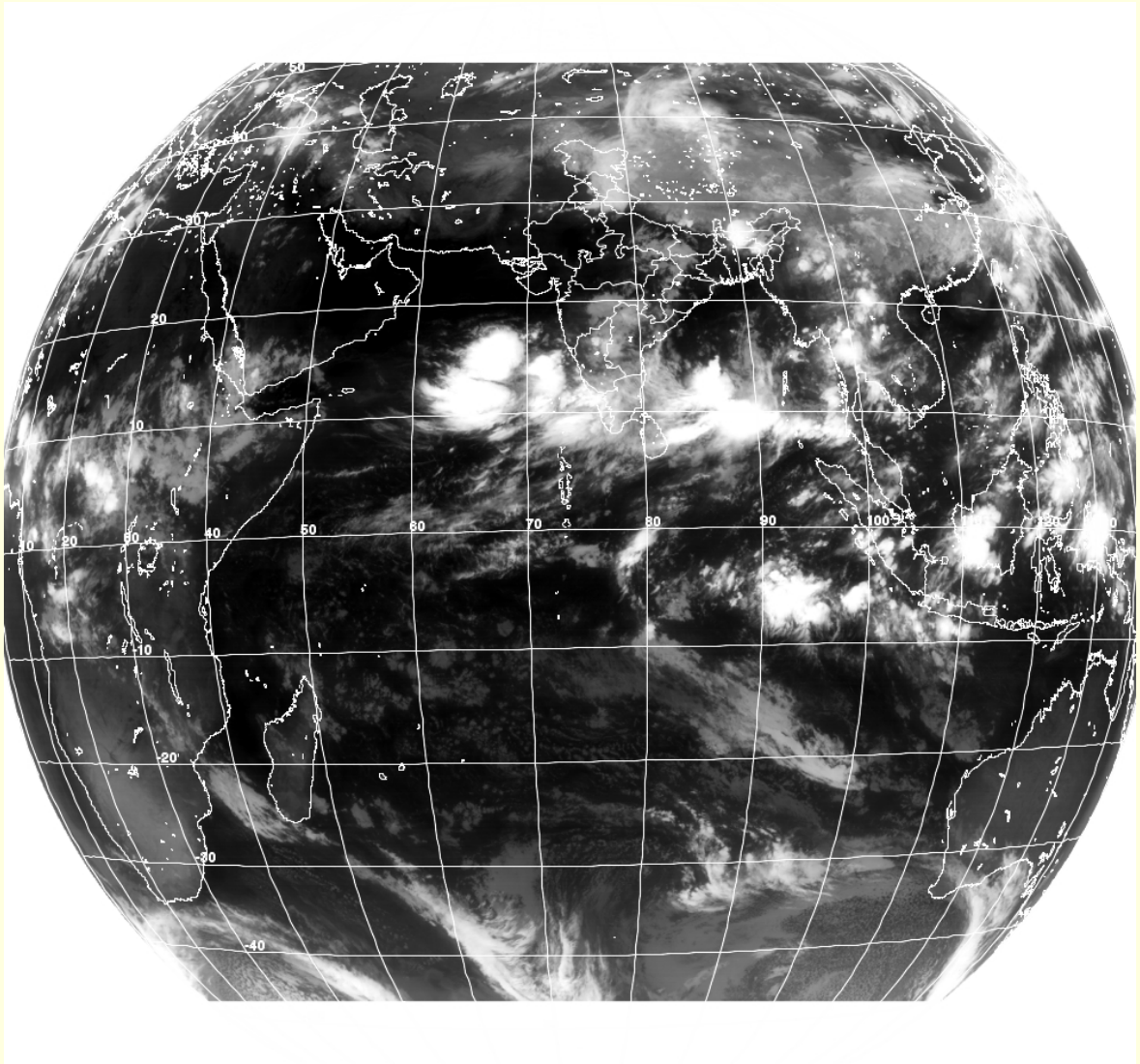


The second hypothesis: seasonal variation of ITCZ



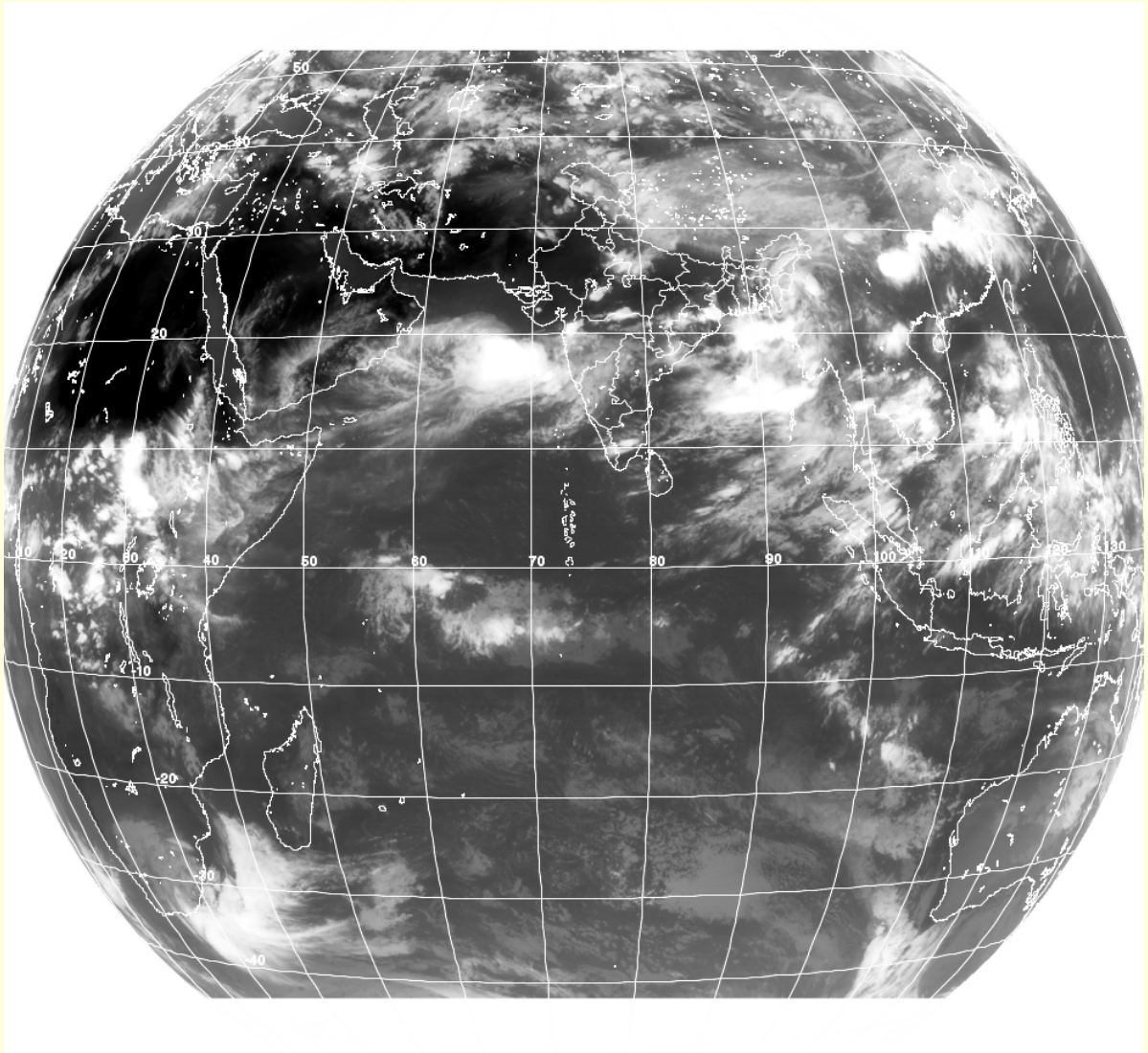
28 May 2011

The second hypothesis: seasonal variation of ITCZ



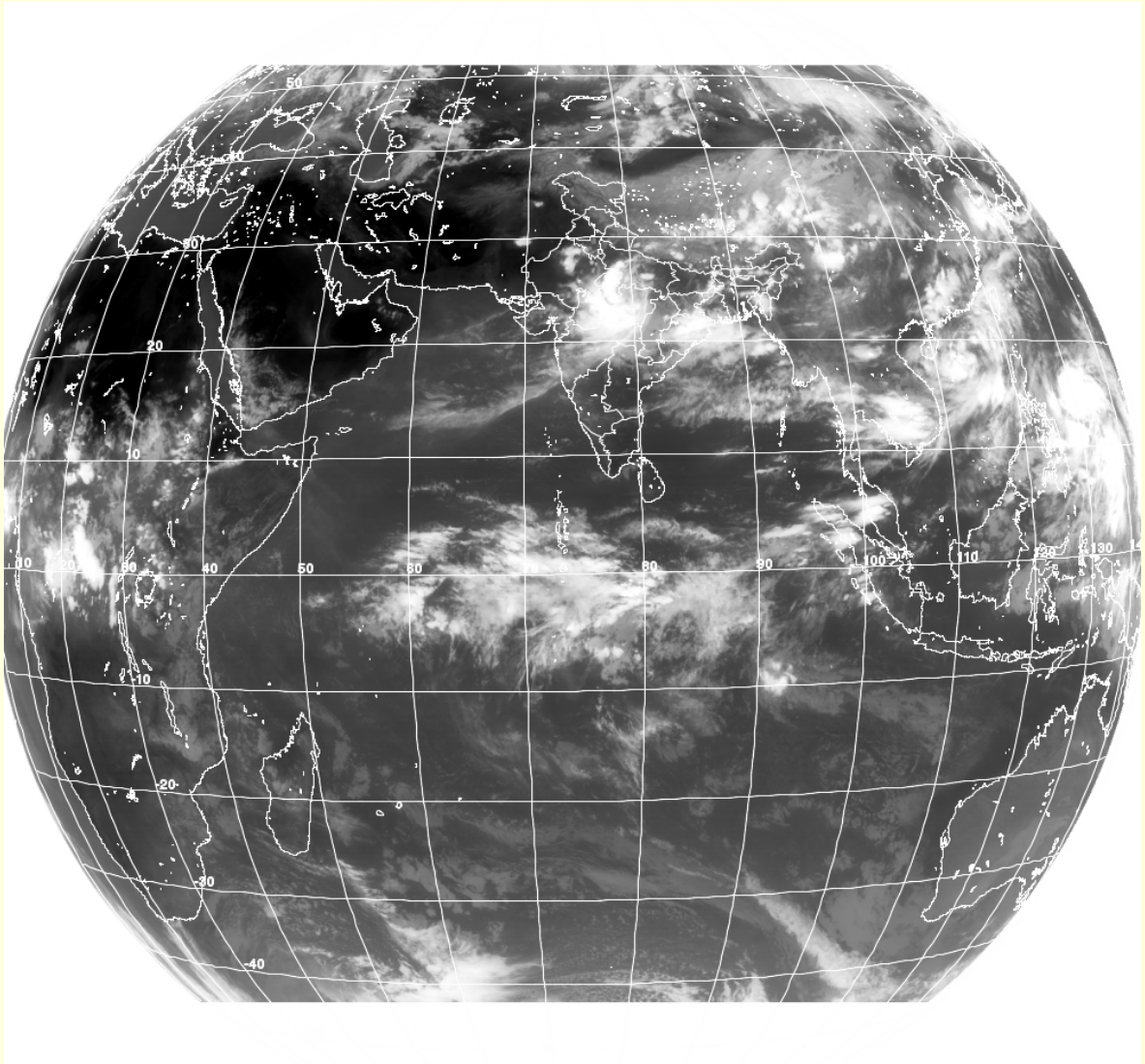
2 June 2011

The second hypothesis: seasonal variation of ITCZ



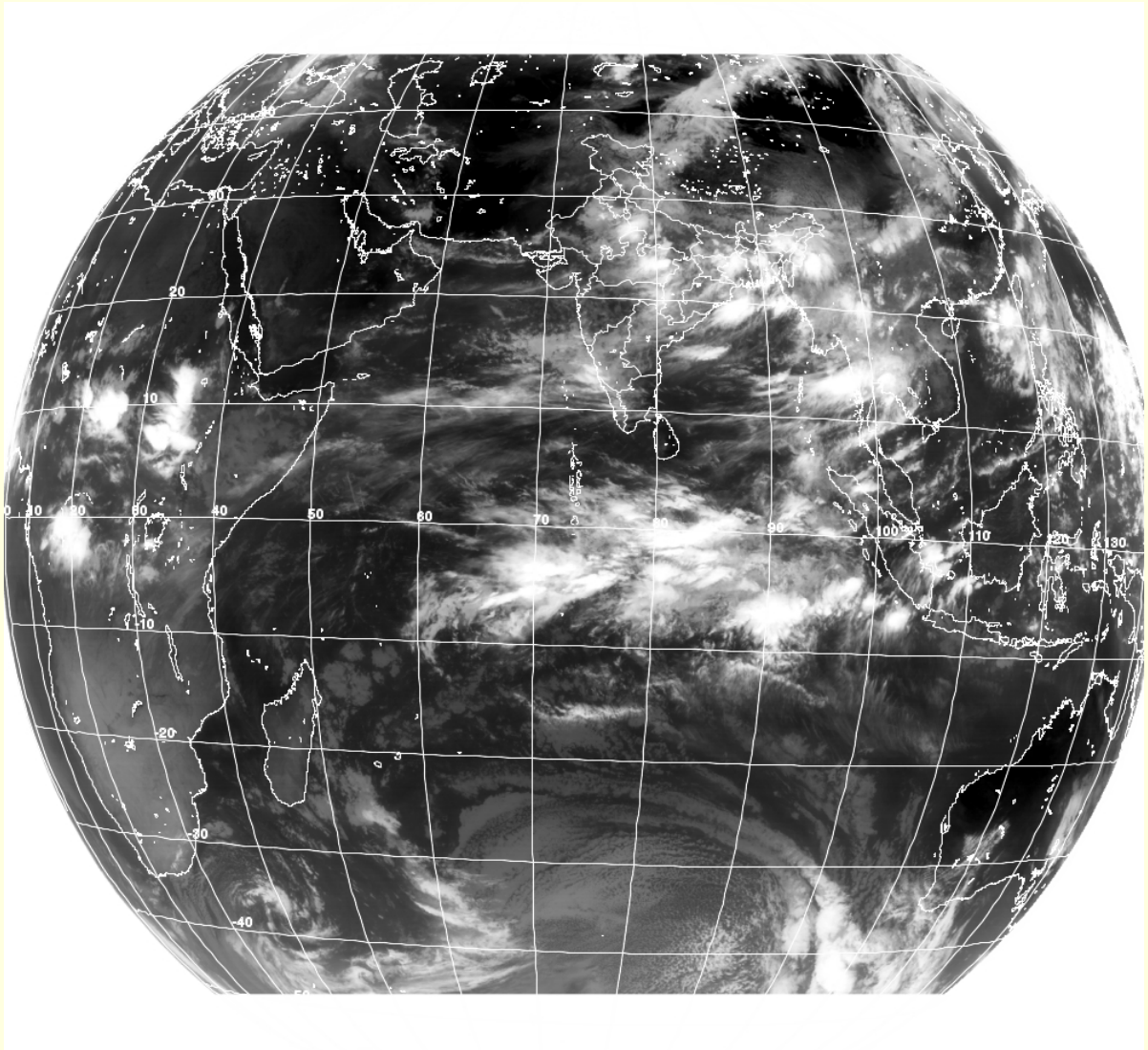
9 June 2011

The second hypothesis: seasonal variation of ITCZ



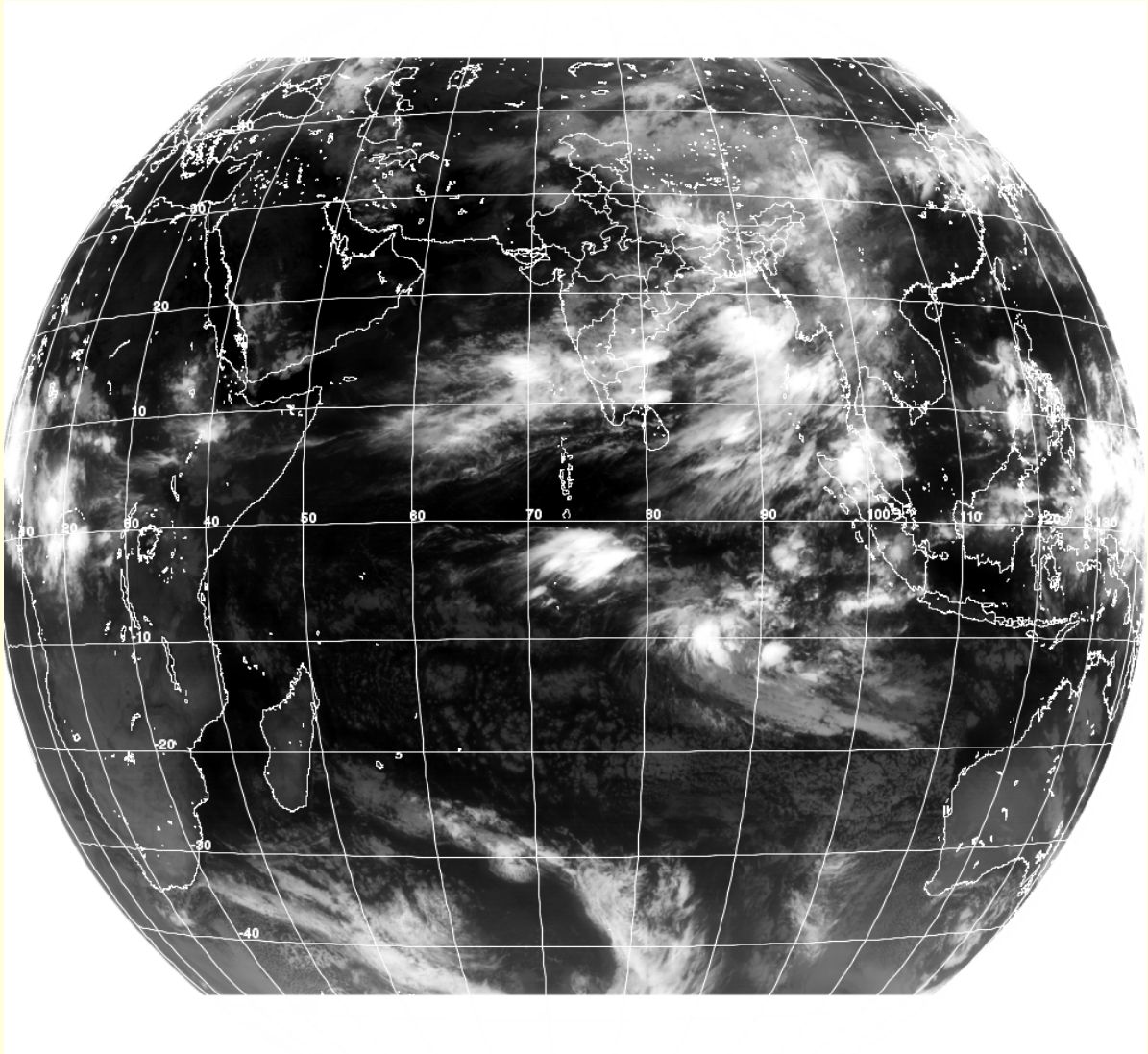
21 June 2011

The second hypothesis: seasonal variation of ITCZ



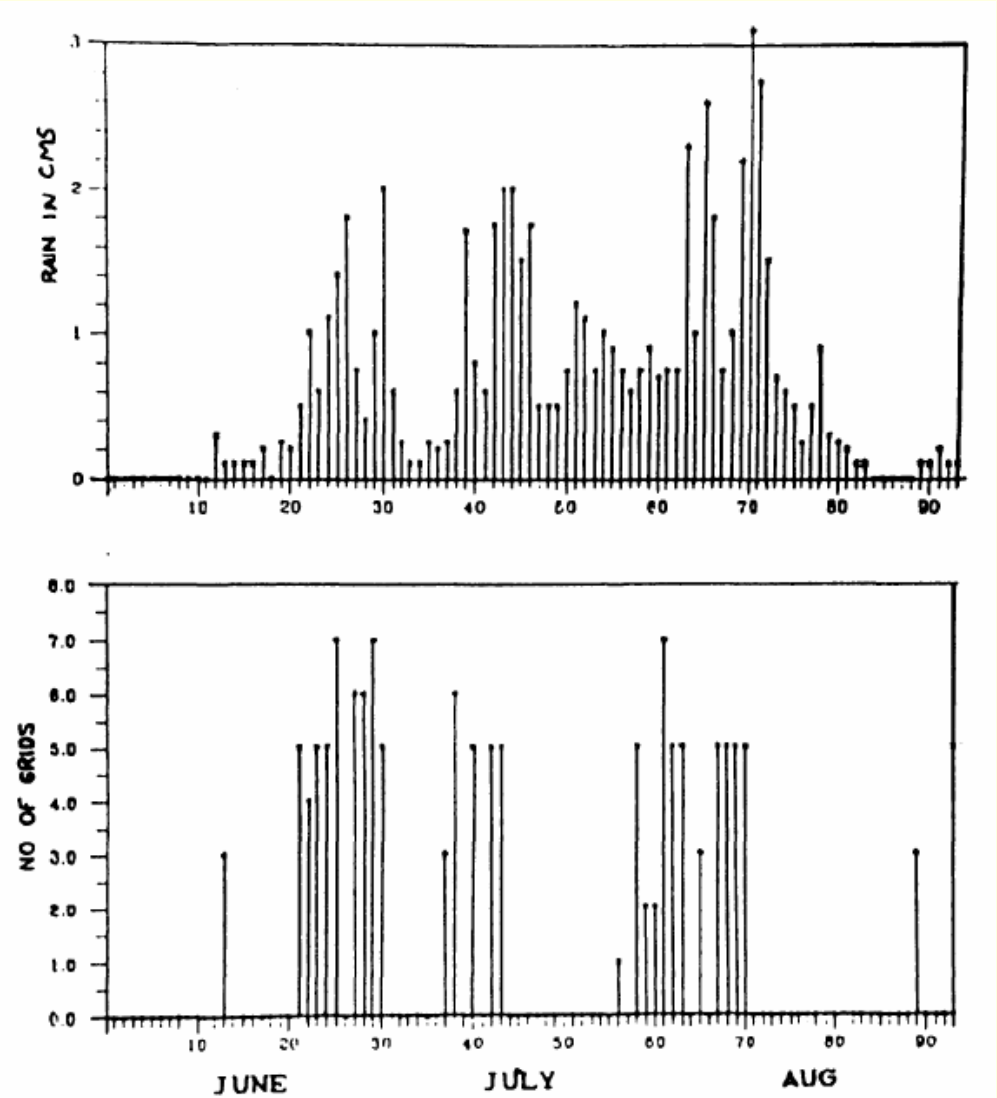
28 June 2011

The second hypothesis: seasonal variation of ITCZ



3 July 2011

The second hypothesis: seasonal variation of ITCZ



- top: ISMR during 1979
- bottom: fractional area around 80E covered by ITCZ clouds

The second hypothesis: seasonal variation of ITCZ

- Some "simple" PDE models (between 6-10 PDE in two spatial variables, zonally averaged), studied in a series of papers by Webster, Chou, Gadgil, Srinivasan, Nanjundiah, et al. to capture the northward propagation of TCZ
- "meridional variation in the cumulus heating" due to
- " meridional variation in the convective instability" which is due to
- "meridional variation of the surface temperature"
- Again, the role of heating and convection is crucial.
- This leads to the simple model of Gill (QJRMS, 1980, v.106, p.447) "to elucidate some features of the response of the tropical atmosphere to diabatic heating

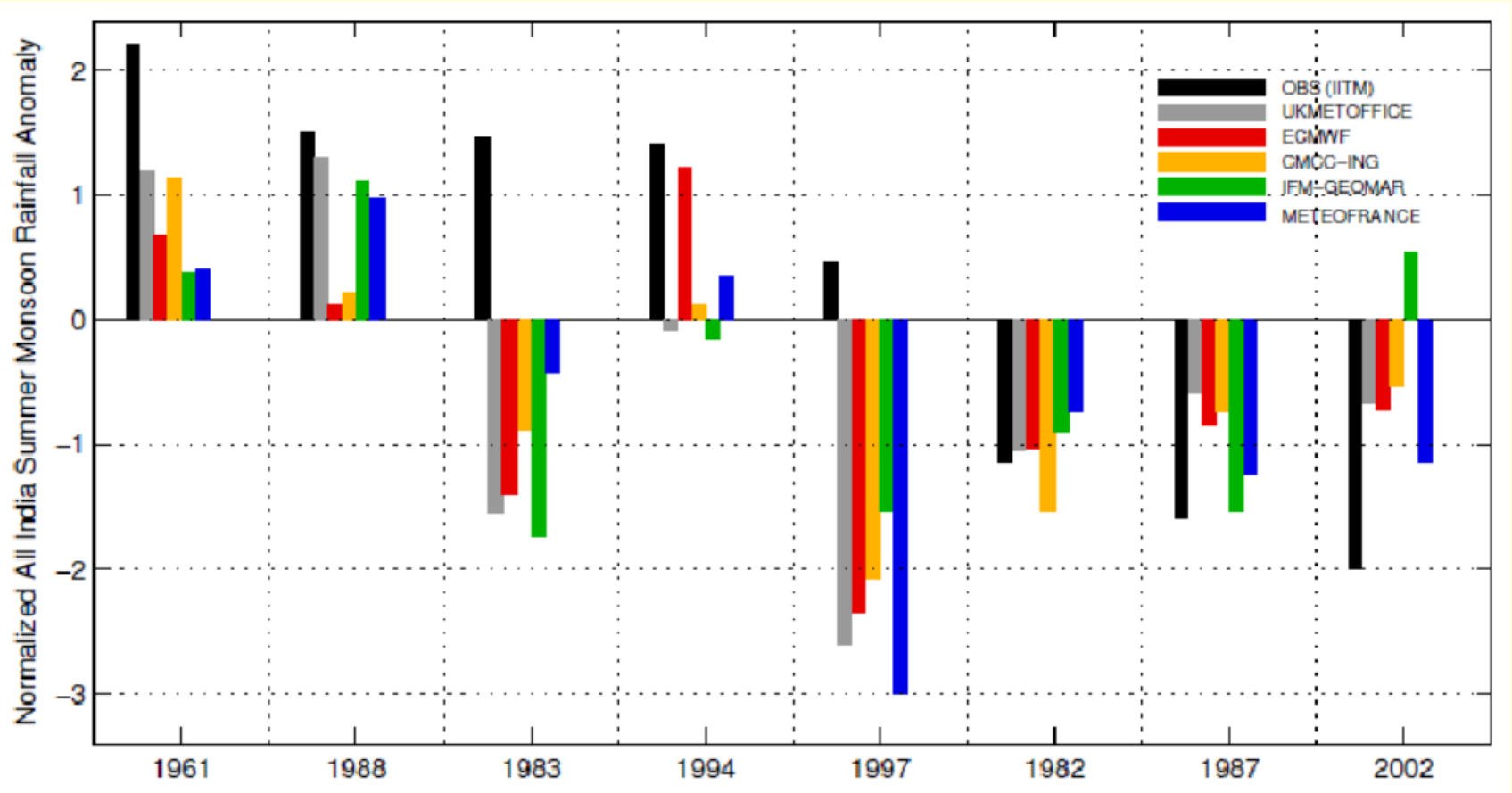
But convection is one of the most uncertain parts of general circulation models...

How well do the general circulation models predict the monsoon?

Poorly! obligatory quote about predicting the future...

- Wang et al. (2004) J. Climate, v.17, p.803 or Gadgil and Srinivasan, Curr. Sci, v.100, p.343: Atmospheric GCM with specified SST fails to predict monsoon, typical quotes: "as expected from earlier studies, that none of the models were able to simulate the correct sign of the anomaly of the Indian summer monsoon rainfall for all the years"
- Jiang et al. doi:10.1175/JCLI-D-12-00437: coupled models may perform better, e.g., CFSv2

"Synchronization of models"?!



- All models may predict it: 1982, '87, '61
- Or none do: 1983, '97

Nanjundiah et al. 2013, curr.sci., v.104, p.1380

Summary so far

- Interannual variability is small but has significant impact
- Relations to ENSO and EQUINOO is important for understanding this variability - very little mathematical modeling of this aspect
- There are complex spatio-temporal variations within a season (intraseasonal variability)
- Heating and convection (and the ITCZ) play an important role in understanding this variability. But I will not talk about the Gill, Lindzen-Nigam, Neelin models of this aspect, nor about the relations to Madden-Julian oscillations (MJO).
- General circulation models (GCMs) have difficulty in capturing the interannual as well as the intraseasonal variability of monsoon
- Generally, balance of energy and moisture affects the dynamics greatly

Data based models attempt to understand the spatio-temporal patterns, hopefully leading to physical insights that can be used to improve the GCMs.

Outline

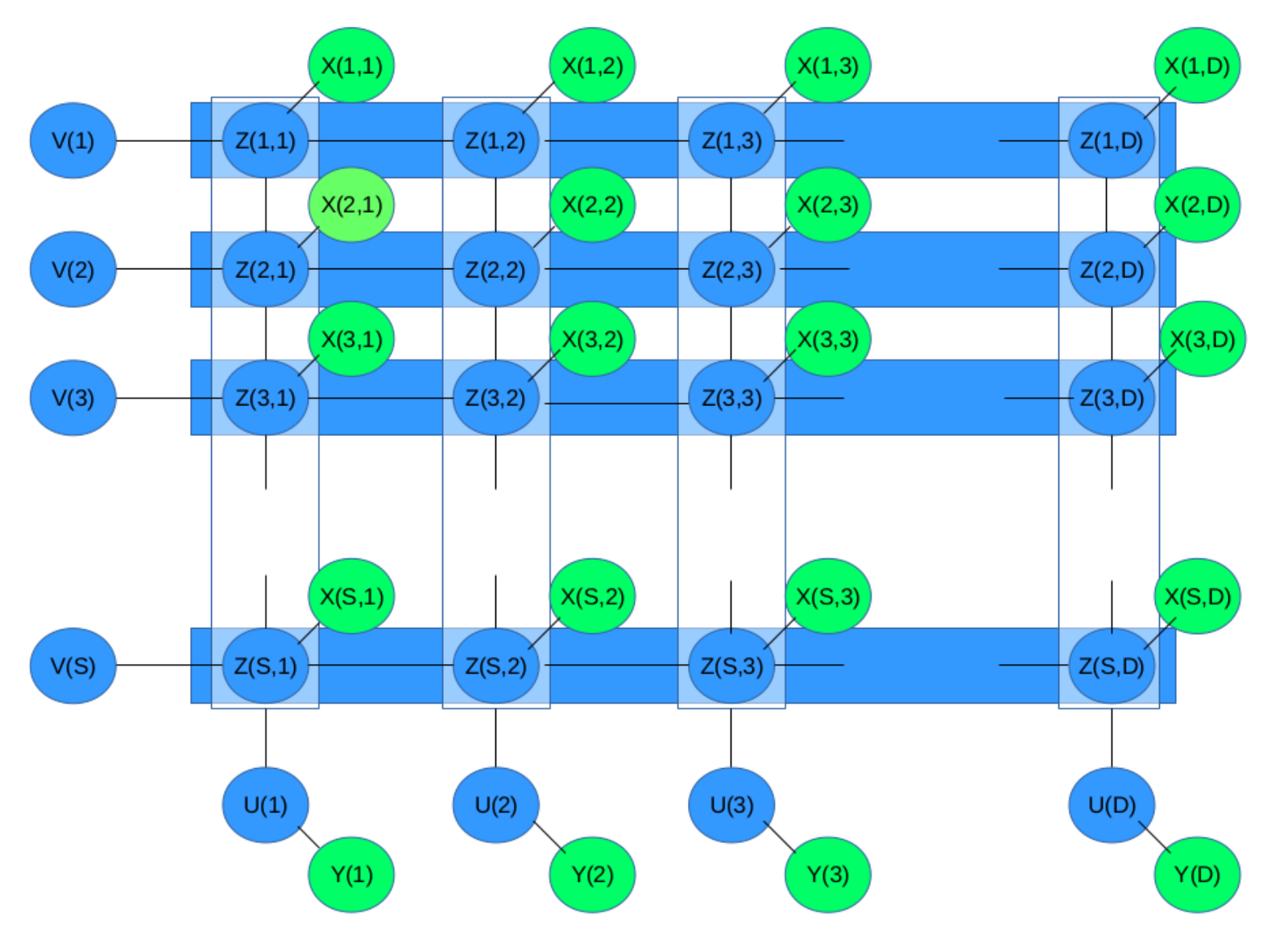
Monsoon: a brief description

Markov random field (MRF) model

Results: prominent spatial patterns

Discussion

MRF: a network random variables at nodes and probability distributions on the edges



Nodes: discrete and continuous random variables

- $Z(s,t) \in \{0,1\}$ indicating low and high rainfall states at location s on day t
- $U(t) \in \{1,\ldots,L\}$: integer valued; indicates the membership of the day t to a cluster of days with cluster label $U(t)$
- $V(s) \in \{1,\ldots,K\}$: integer valued; indicates the membership of the location s to a cluster of locations with cluster label $V(s)$
- $X(s,t)$: real-valued continuous random variable indicating the rainfall at location s on day t

We study the conditional distribution $p(Z, U, V|X = x)$

- MRF model defined by the dependency structure between the nodes as given by the edges of the graph
- Edge potentials associated with the edges define the joint probability distribution $p(Z, U, V, X)$
- The available rainfall data $x(s, t)$ for $s = 1, \ldots, S$ and $t = 1, \ldots, D$ is a specific realization $X = x$ on which to condition the probability distribution of other three variables Z, U, V
- The central inference step involves sampling from the conditional distribution $p(Z, U, V|X = x)$. We use Gibbs sampling algorithm.

Patterns are obtained by averaging over clusters

$$\begin{aligned} \phi_u(s) &= \text{mean}_t\left(x(s,t) : U(t) = u\right), \\ \phi_u^d(s) &= \text{mode}_t\left(Z(s,t) : U(t) = u\right) \end{aligned}$$

These S-dimensional vectors are the spatial patterns.

We study the conditional distribution $p(Z, U, V|X = x)$

- MRF model defined by the dependency structure between the nodes as given by the edges of the graph
- Edge potentials associated with the edges define the joint probability distribution $p(Z, U, V, X)$
- The available rainfall data $x(s, t)$ for $s = 1, \ldots, S$ and $t = 1, \ldots, D$ is a specific realization $X = x$ on which to condition the probability distribution of other three variables Z, U, V
- The central inference step involves sampling from the conditional distribution $p(Z, U, V|X = x)$. We use Gibbs sampling algorithm.

Patterns are obtained by averaging over clusters

$$\begin{aligned} \theta_u(t) &= \text{mean}_s\left(x(s,t) : V(s) = v\right), \\ \theta_u^d(t) &= \text{mode}_s\left(Z(s,t) : V(s) = v\right) \end{aligned}$$

These D-dimensional vectors are the temporal patterns.

"Edge potentials" define an MRF

In a undirected graph (V, E)

- If $v, u \in V$ are connected by an edge, then define a "edge potential" $\psi(u, v)$, which is a probability distribution
- The probability distribution of the nodes is just a product of all edge potentials: $p(V) \propto \prod_{e \in E} \psi(e)$

Main idea: the edge potentials can be used to "encode" domain knowledge: for example

- for the variables Z: threshold for high/low rainfall in terms of the mean of the edge potential
- for clustering variables U: how well the spatial patterns align with the pattern for each day

Edge potentials define "inter-dependency" of these variables

- Edges between Z and X are Gamma distributions:

$$\psi_{DZ}(Z(s,t)=z, X(s,t)) \;=\; (X(s,t))^{\alpha_{sz}-1}\exp\left(-\beta_{sz}X(s,t)\right) \quad (1)$$

- The parameters α, β are inferred as part of the modeling process
- Edges between U, V and Z variables are exponential distributions:

$$\begin{aligned}\psi_{SS}(Z(s,t), U(t)) &= \exp\left(\eta\, \mathbb{1}_{\{Z(s,t)=\phi^d(s,U(t))\}}\right),\\ \psi_{ST}(Z(s,t), V(s)) &= \exp\left(\zeta\, \mathbb{1}_{\{Z(s,t)=\theta^d(V(s),t)\}}\right).\end{aligned}$$

- The parameters η and ζ are "control parameters" in the model
- The edges between the Z-variables at different spatio-temporal locations are used to "control" the spatial coherence of the patterns.

Summary so far

MRF model consisting of:

- Discrete random variables Z, U, V, in order to obtain a "coarse" picture of the monsoon rainfall
- Probabilistic model to incorporate "domain knowledge" in terms of probability distributions for these variables
- Inference in terms of conditional distribution conditioned on observed rainfall data

Main aims of the MRF model

- Clustering of locations and of days, in order to identify
- Dominant patterns in monsoon rainfall data ("model reduction" analogous to techniques such as EOF)

Outline

Monsoon: a brief description

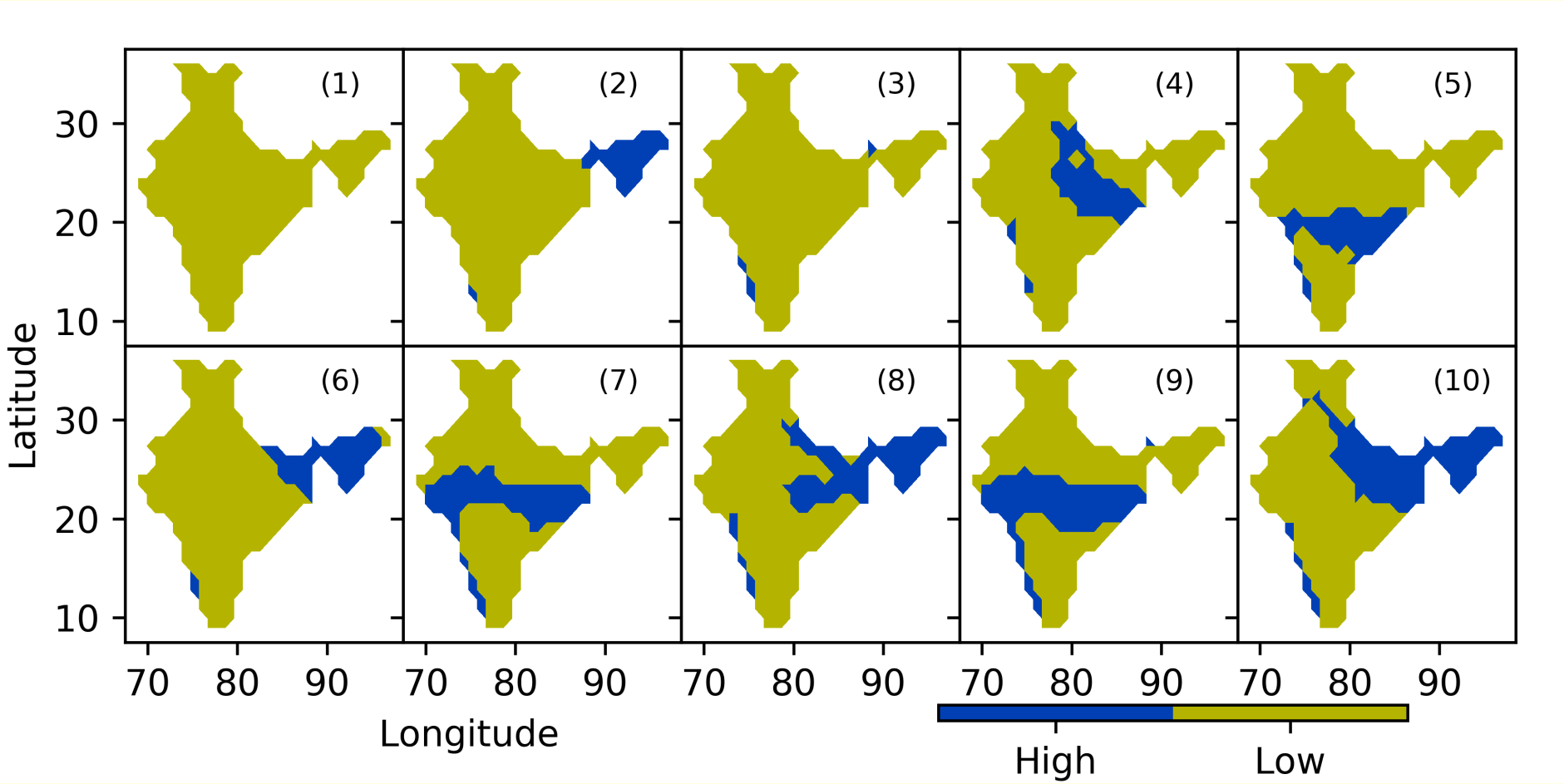
Markov random field (MRF) model

Results: prominent spatial patterns

Discussion

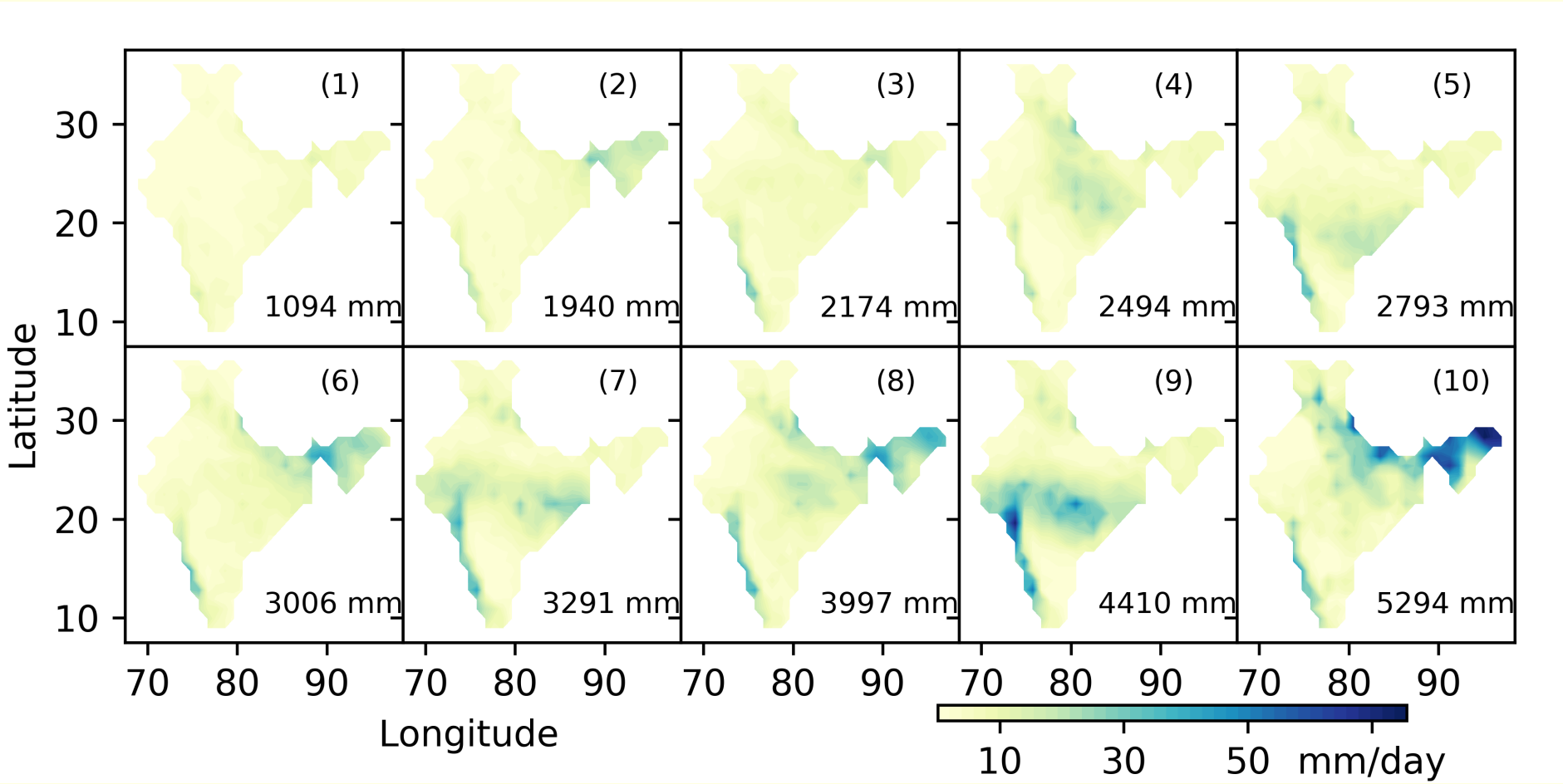
We find 10 prominent patterns

Discrete variable Z



We find 10 prominent patterns

Rainfall data

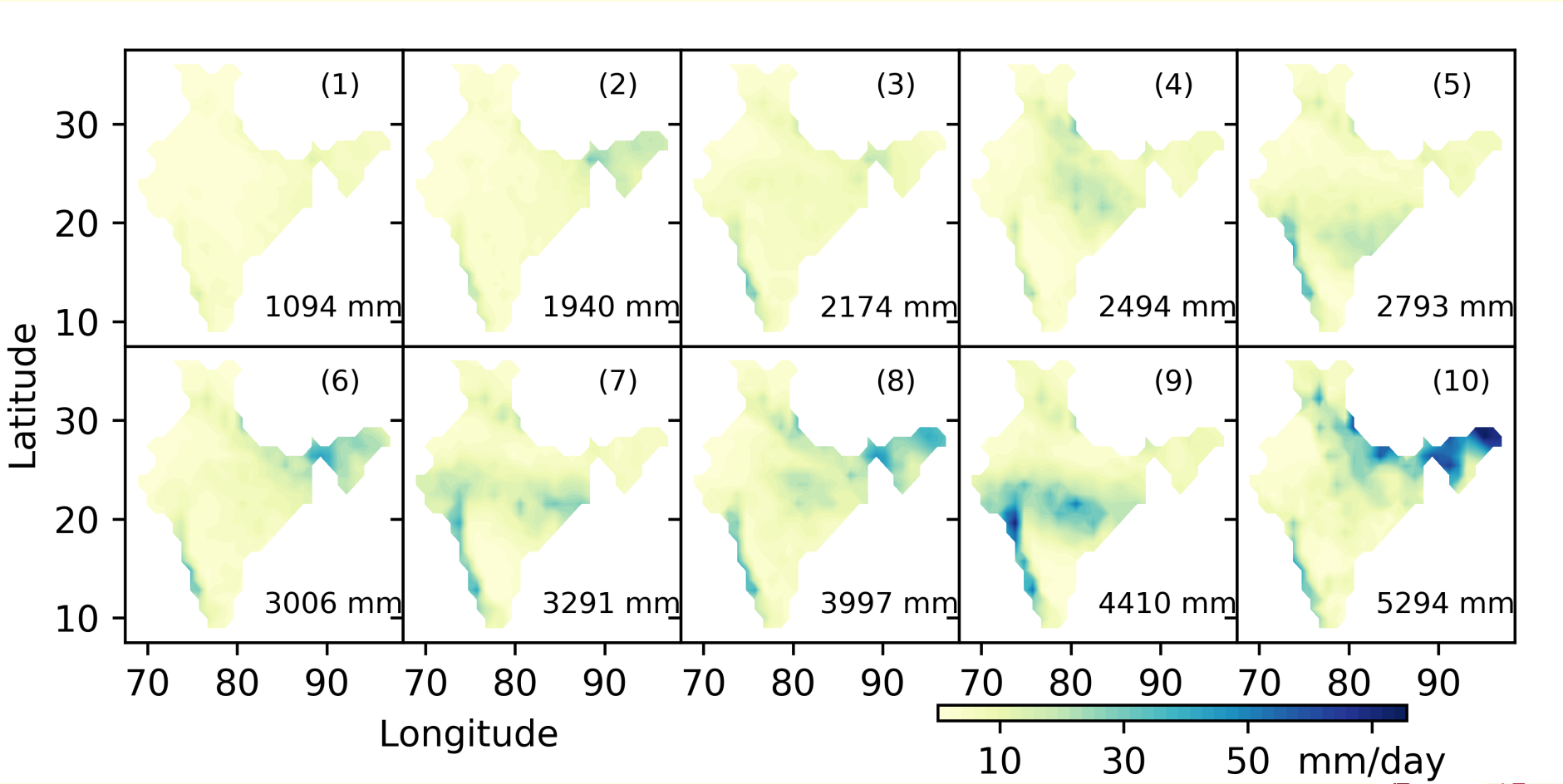


Other methods for clustering / pattern

- K-means and spectral clustering: two commonly used algorithms that find clusters in the "data space" (i.e., directly working with the rainfall data $x(s, t)$)
- Again, for each cluster, we can associate spatial patterns
- EOF: finding the most significant singular vectors to represent the data: naturally gives patterns in data, but not clustering

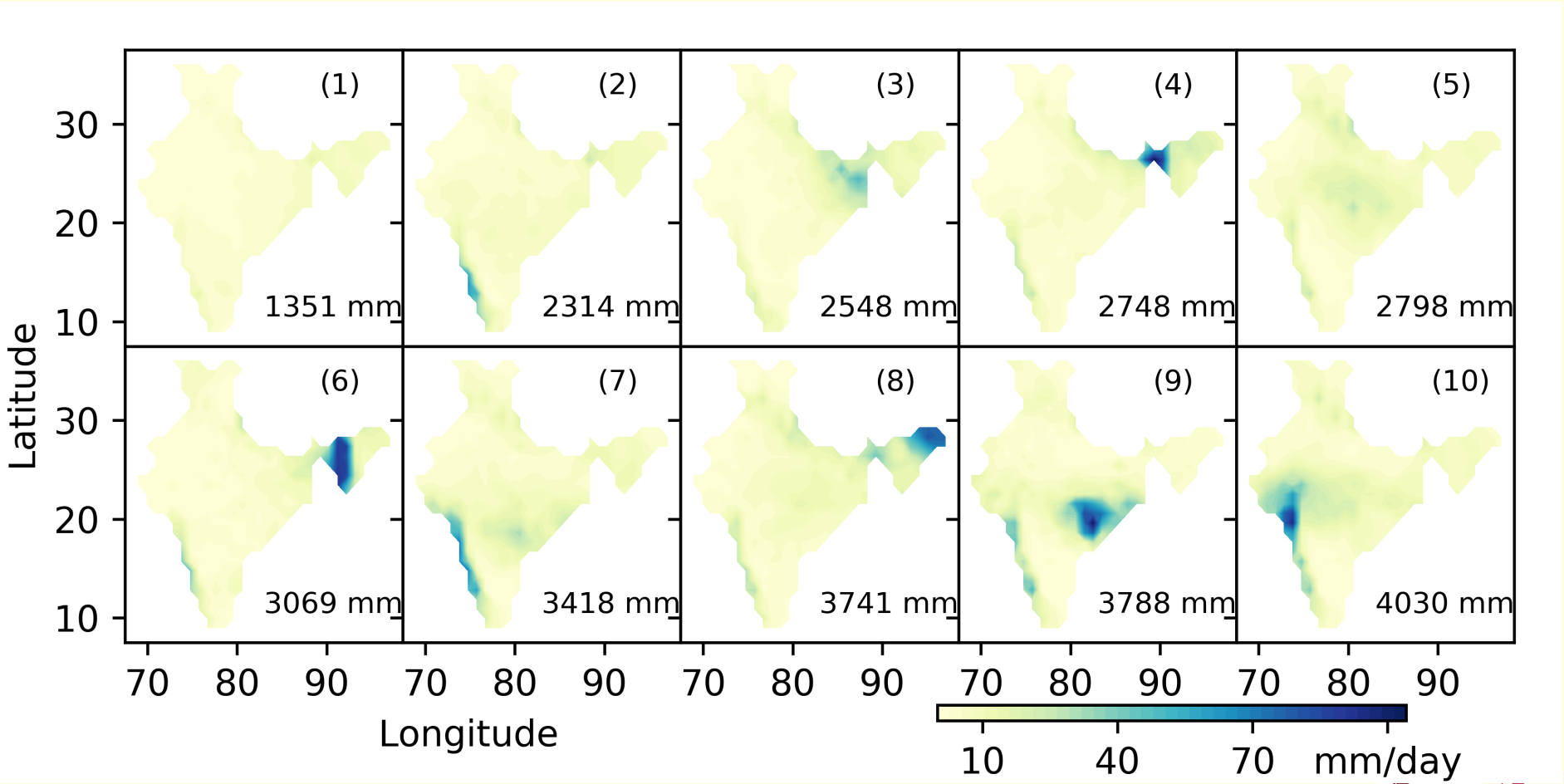
Patterns obtained by MRF are more spatially coherent and more representative

Ten patterns from MRF



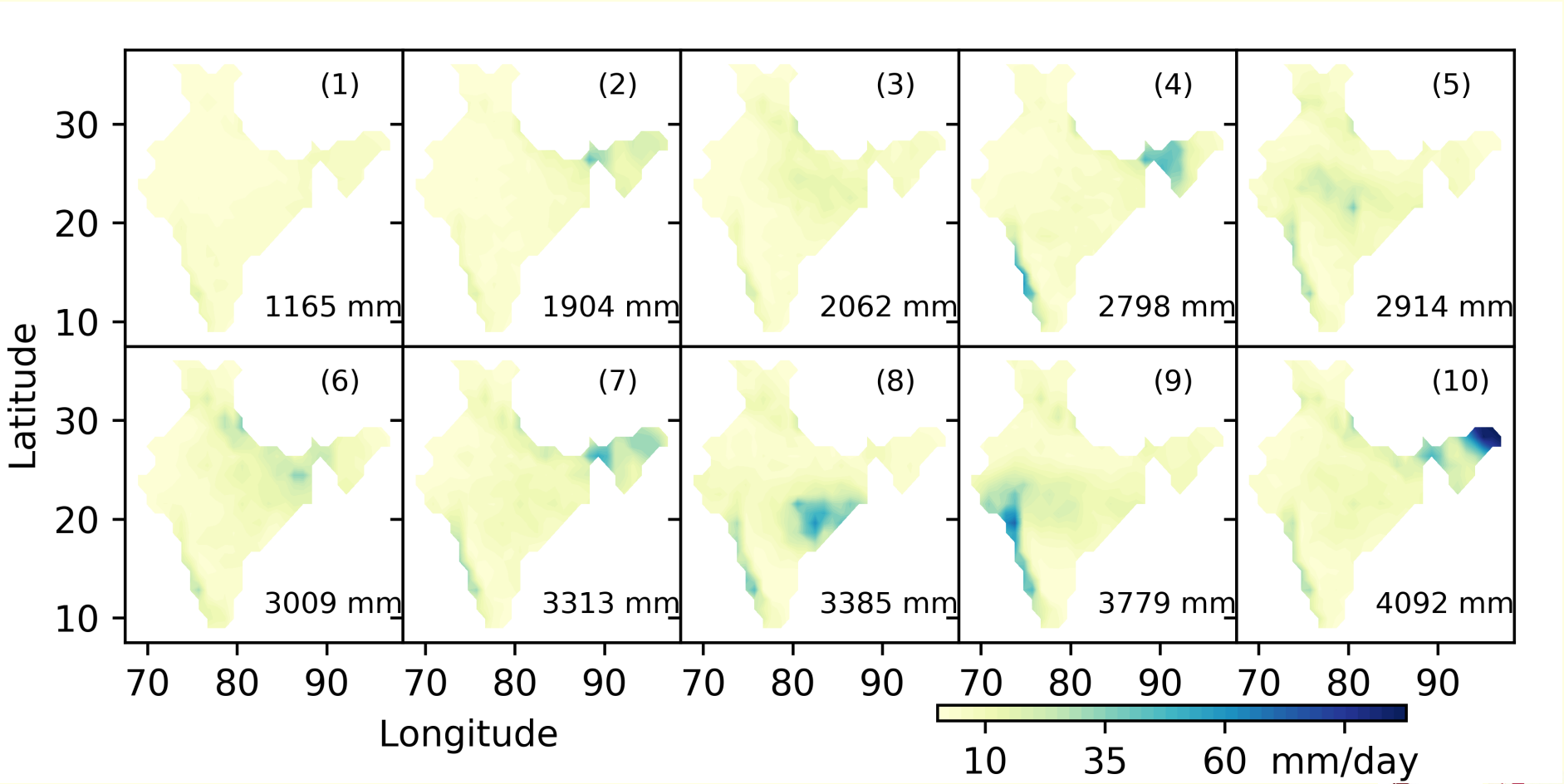
Patterns obtained by MRF are more spatially coherent and more representative

Ten patterns from K-means



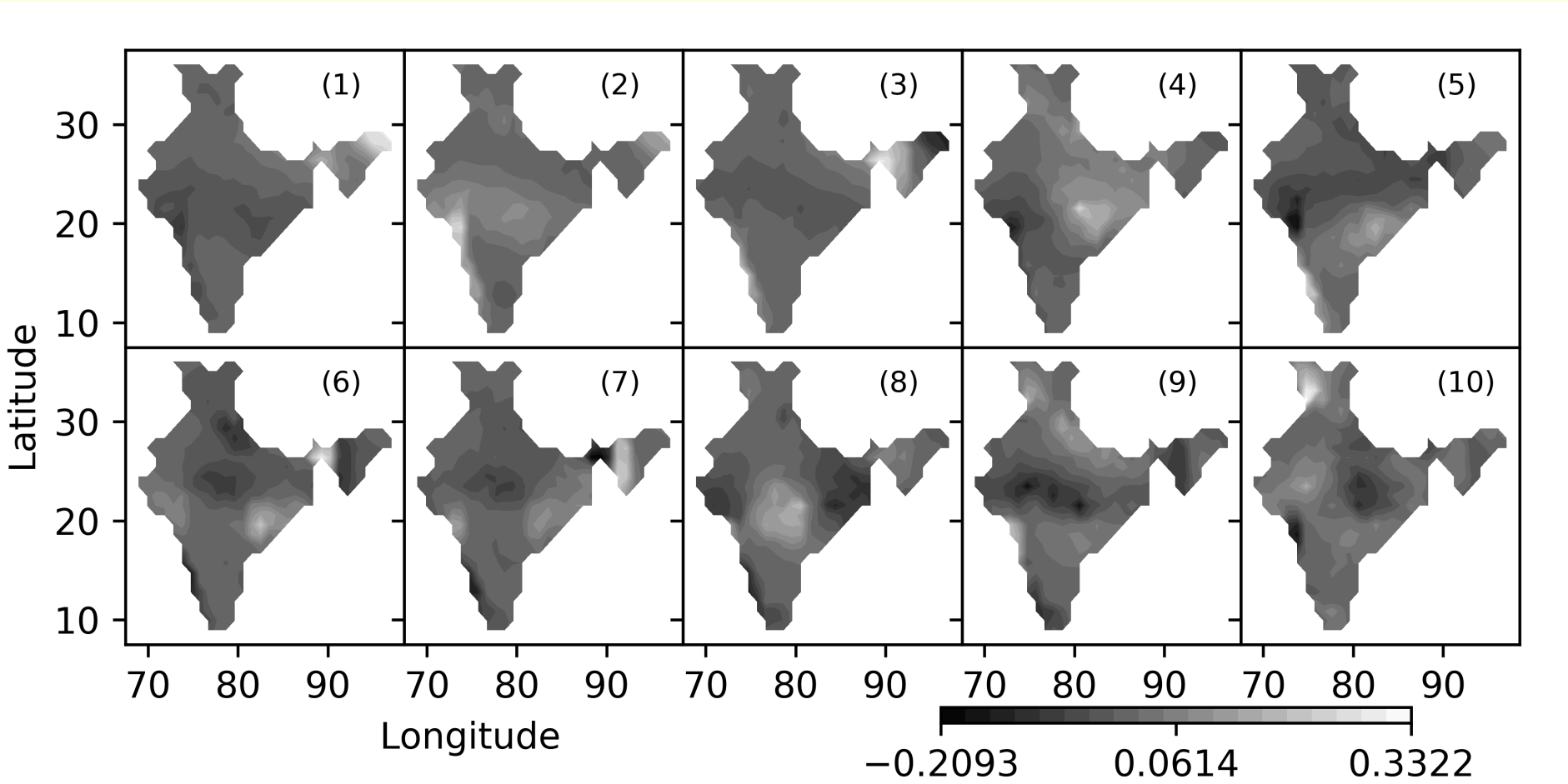
Patterns obtained by MRF are more spatially coherent and more representative

Ten patterns from spectral clustering



Patterns obtained by MRF are more spatially coherent and more representative

Ten patterns from EOF



MRF patterns are representative and coherent

η	#PC				PC coverage (average)				std(Y)		
(#clusters)	MRF	KM	SP1	SP2	MRF	KM	SP1	SP2	MRF	KM	SP1
5 (146)	**11**	26	38	30	**556** (**50.5**)	516 (19.8)	514 (13.5)	378 (12.6)	**1.06**	1.28	1.5
7 (65)	**11**	26	43	44	786 (**71.5**)	735 (28.3)	**816** (19.0)	800 (18.2)	**1.07**	1.73	1.77
8 (36)	**10**	20	34	34	866 (**86.6**)	862 (43.1)	**953** (28.0)	928 (27.3)	**1.06**	1.86	1.76
9 (24)	**10**	18	22	23	938 (**93.8**)	951 (52.8)	**953** (43.3)	944 (41.0)	**1.05**	2.08	1.85
10 (15)	**11**	16	15	15	966 (**87.8**)	965 (60.3)	**976** (65.1)	**976** (65.1)	**1.22**	2.27	1.87

Table 3: Comparison of daily cluster properties, by varying the number of clusters through η parameter of the proposed model. #PC denotes number of prominent clusters (spanning at least 5 years), and PC coverage denotes number of days (out of 976) assigned to the prominent clusters, and the number in parenthesis gives the average number of days per prominent cluster. The last columns give the standard deviation of the aggregate daily rainfall for days assigned to a cluster. The best performing value is highlighted in bold.

MRF patterns are representative and coherent

$\ell_2(\phi)$			Hamm(ϕ_d)			Agg(ϕ)		
MRF	KMeans	Spect1	MRF	KMeans	Spect2	MRF	KMeans	Spect1
261	263	262	**104**	202	187	**0.49**	0.7	0.75

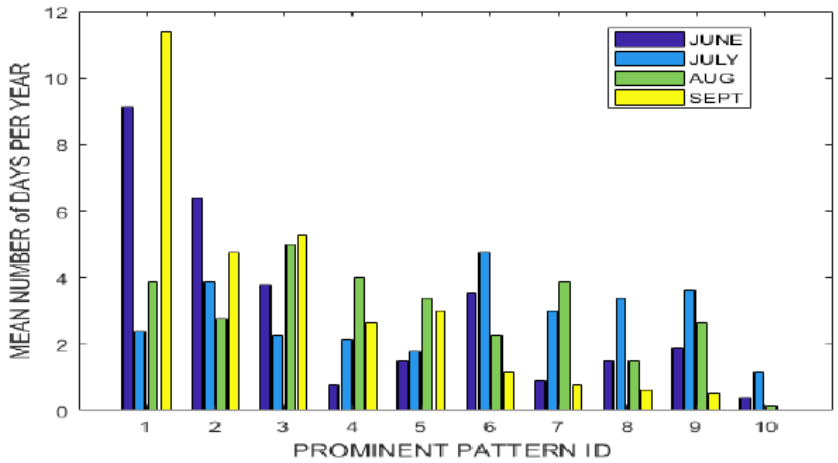
Table 5: Measures of how well the spatial patterns (CRP and CDP) computed over the period 2000-2007 can approximate the daily vectors (DRVs and DDVs) across the period 1901-2011. Three measures are considered: $\ell_2(\phi)$, Hamm(ϕ_d), and Agg(ϕ) are define in equations (14)-(16)

$spch(\phi_d)$			
MRF	KMeans	Spect2	EOF
0.07	0.16	0.14	0.13

Table 6: Measure of spatial coherence of the CDPs discovered by the different methods.

Dynamics of these patterns

Different patterns are associated to periods



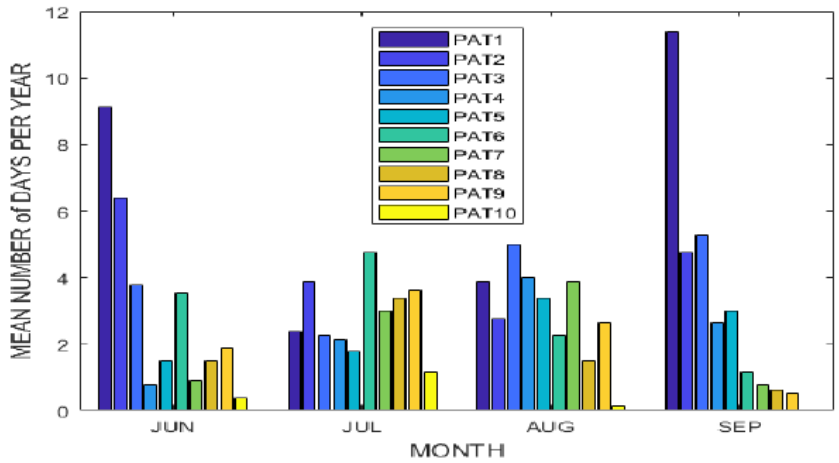


Figure 4: Left: Average number of days under each prominent pattern that belong to the 4 monsoon months (based on the period 2000-2007). Right: Average number of days in each of the 4 monsoon months that were assigned to each prominent pattern (based on the period 2000-2007).

Dynamics of these patterns

We can consider a Markov chain of these patterns

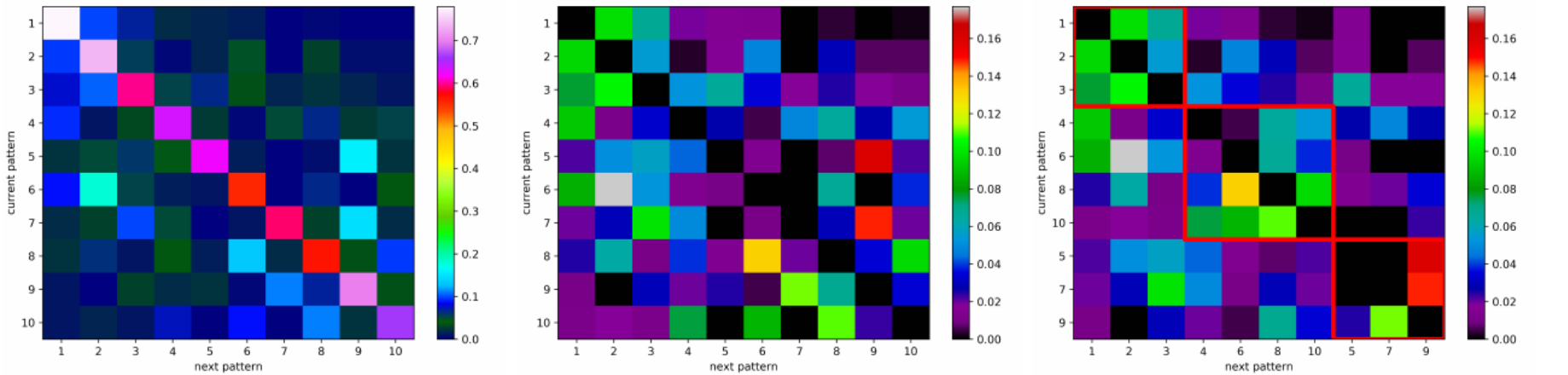
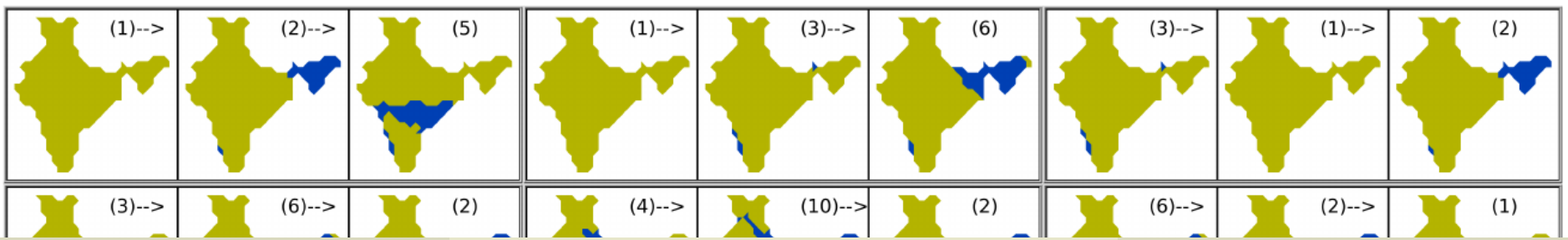


Figure 7: Transition matrix of the patterns between consecutive days. In the left matrix the diagonal elements dominate, indicating strong tendencies of self-transition in each state. In the middle matrix, the diagonal elements have been set to 0 to highlight the transitions other than same-state transitions. In the right matrix, the states have been rearranged according to Families 1,2,3 to highlight the block-diagonal nature of the matrix. Each block along the diagonal represents a family, and we see that intra-family transitions are more frequent than inter-family transitions.

Some transitions appear very frequently



Outline

Monsoon: a brief description

Markov random field (MRF) model

Results: prominent spatial patterns

Discussion

Avenues for further exploration

- Include other variables: some promising results already
- Further study of the Markov dynamics of the patterns
- "Simple" dynamical models that mimic the clustering and patterns obtained from the MRF model of data
- Physical interpretation that may be useful to earth scientists to improve the global circulation models

Thanks:
Infosys Foundation;
Airbus Chair program at ICTS and CAM, TIFR;