Markov random field model for the Indian monsoon rainfall

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

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

14 August 2018
IWCMS, IITM-Pune

Outline

Markov random field (MRF) model

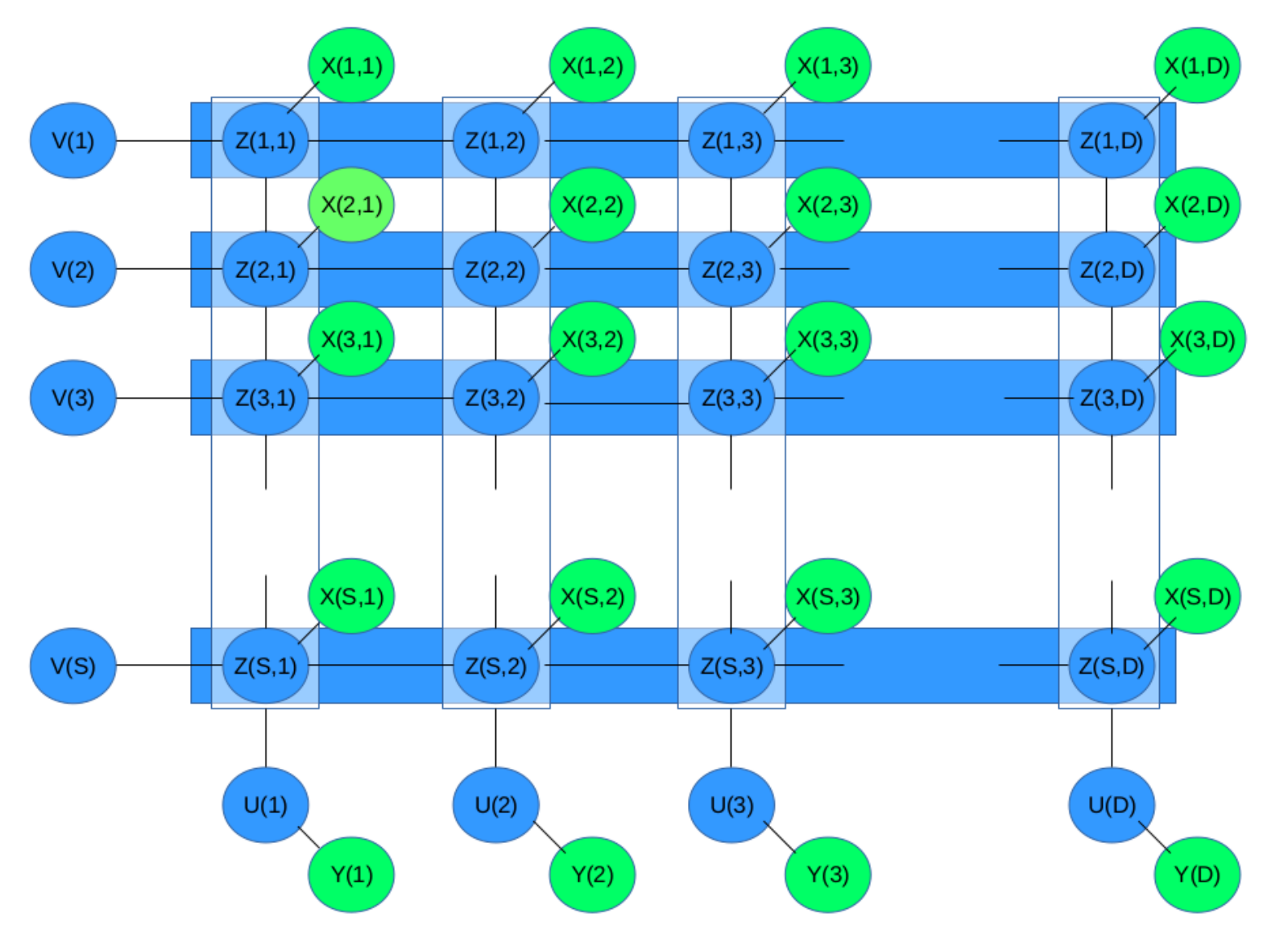
Results: prominent spatial patterns

Discussion

Talk based on arxiv:1805.00414; arxiv:1805.00420

Outline

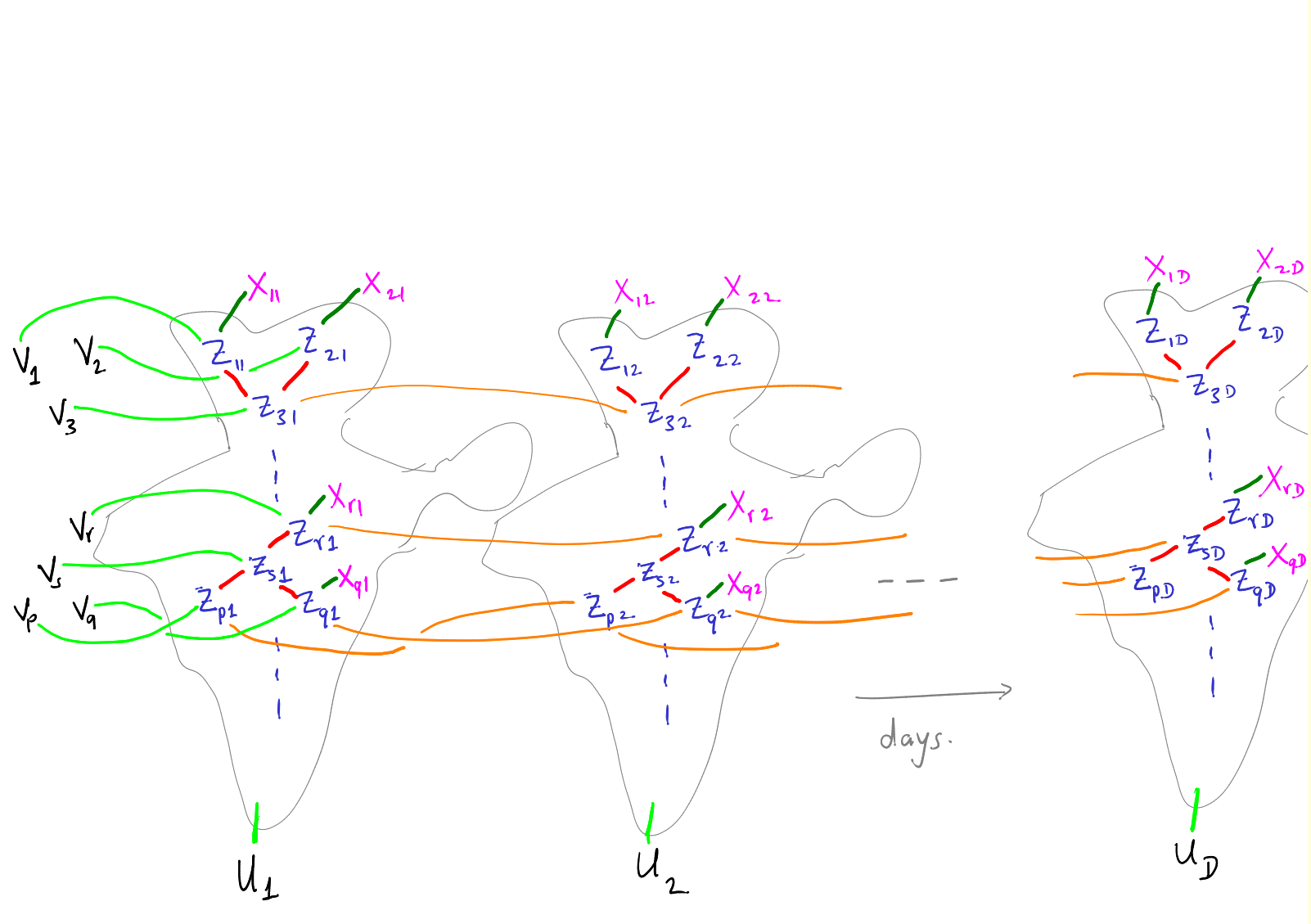
MRF: a network random variables at nodes and probability distributions on the edges



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

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



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 $p(Z, U, V|X = x)$

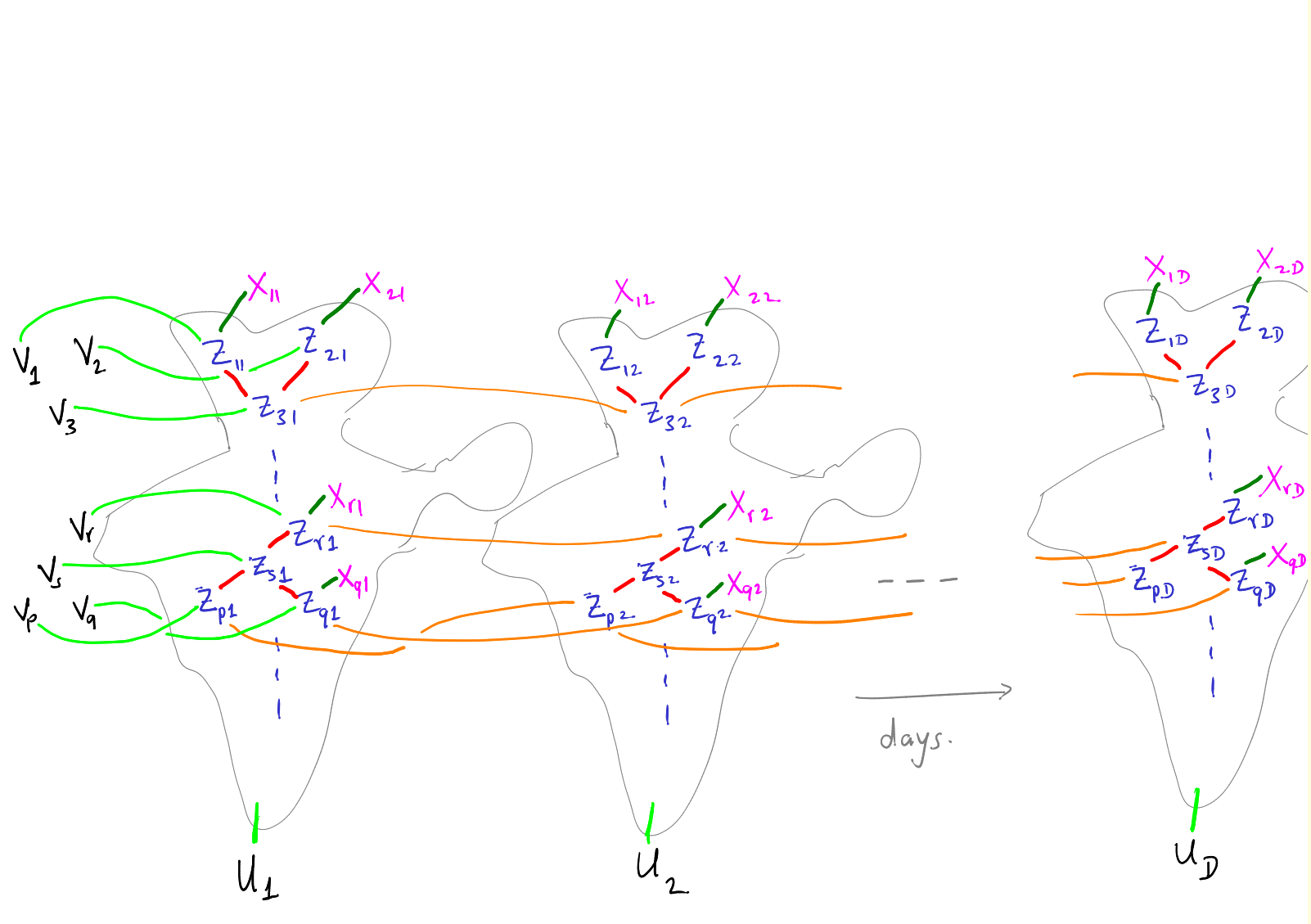
- MRF model defined by the dependency structure between the nodes as given by the edges of the graph
- Edge potentials associated with the edges define the joint probability distribution $p(Z, U, V, X)$
- The available rainfall data $x(s, t)$ for $s = 1, \ldots, S$ and $t = 1, \ldots, D$ is a specific 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

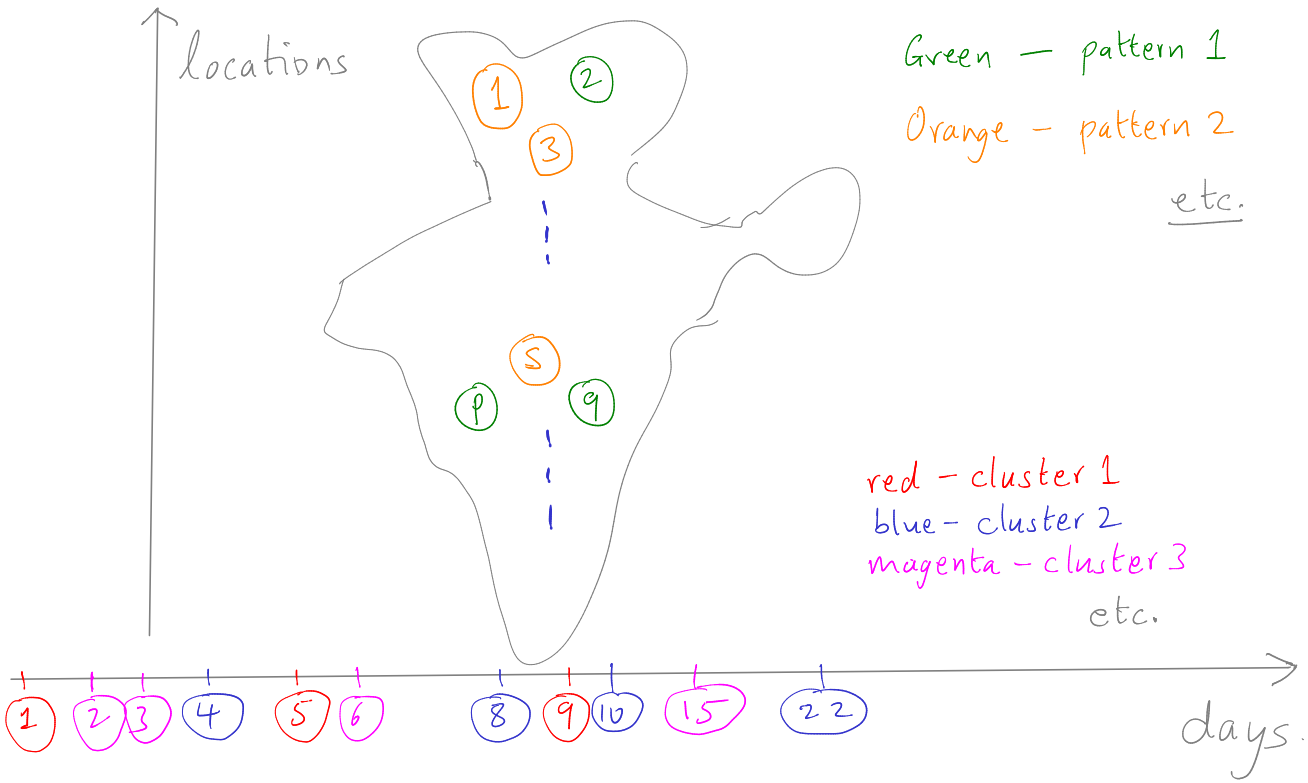
$$\begin{aligned} \theta_u(t) &= \text{mean}_s\left(x(s,t) : V(s) = v\right), \\ \theta_u^d(t) &= \text{mode}_s\left(Z(s,t) : V(s) = v\right) \end{aligned}$$

These D-dimensional vectors are the temporal patterns.

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



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



"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

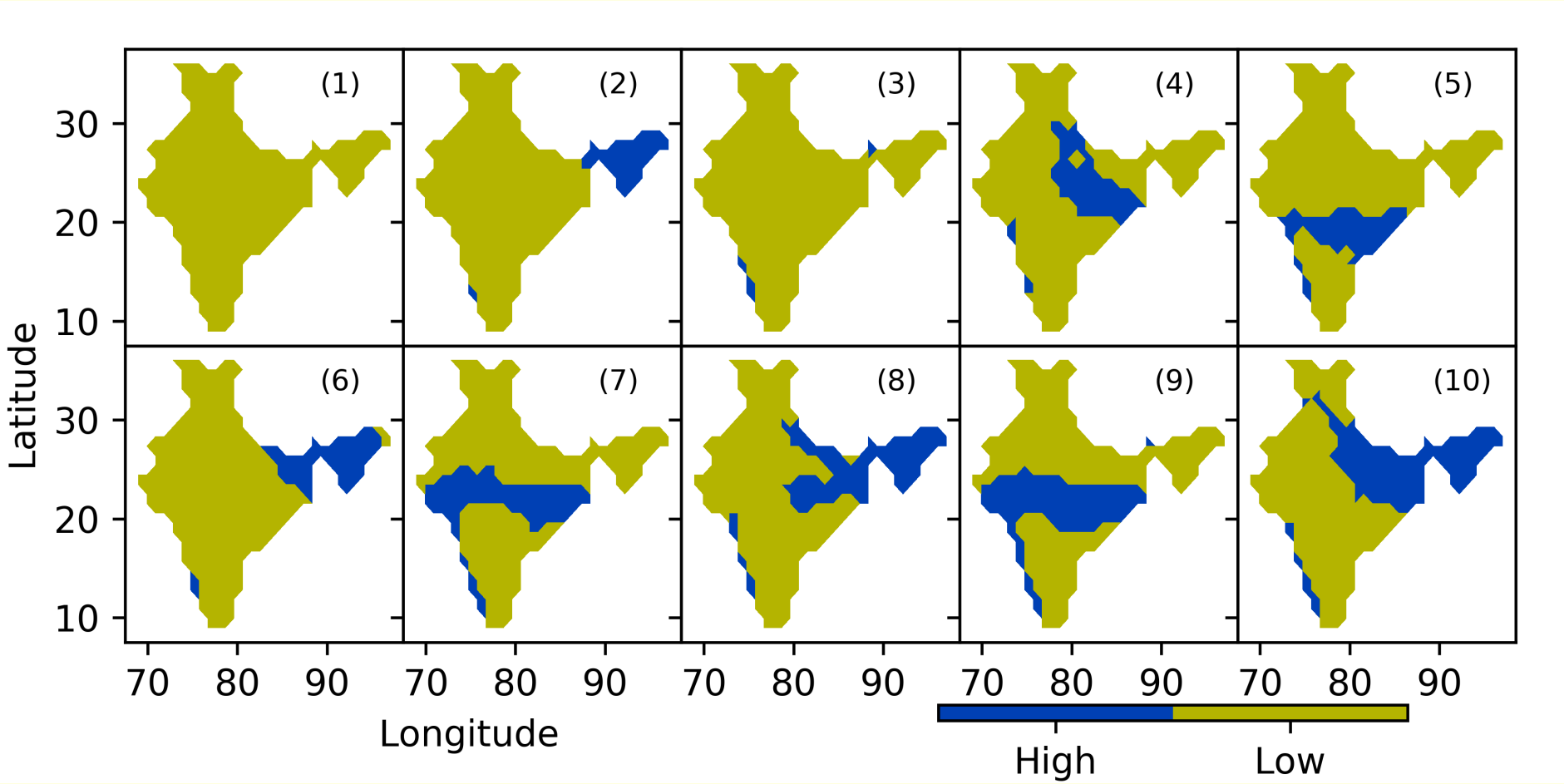
Main aims of the MRF model

- 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

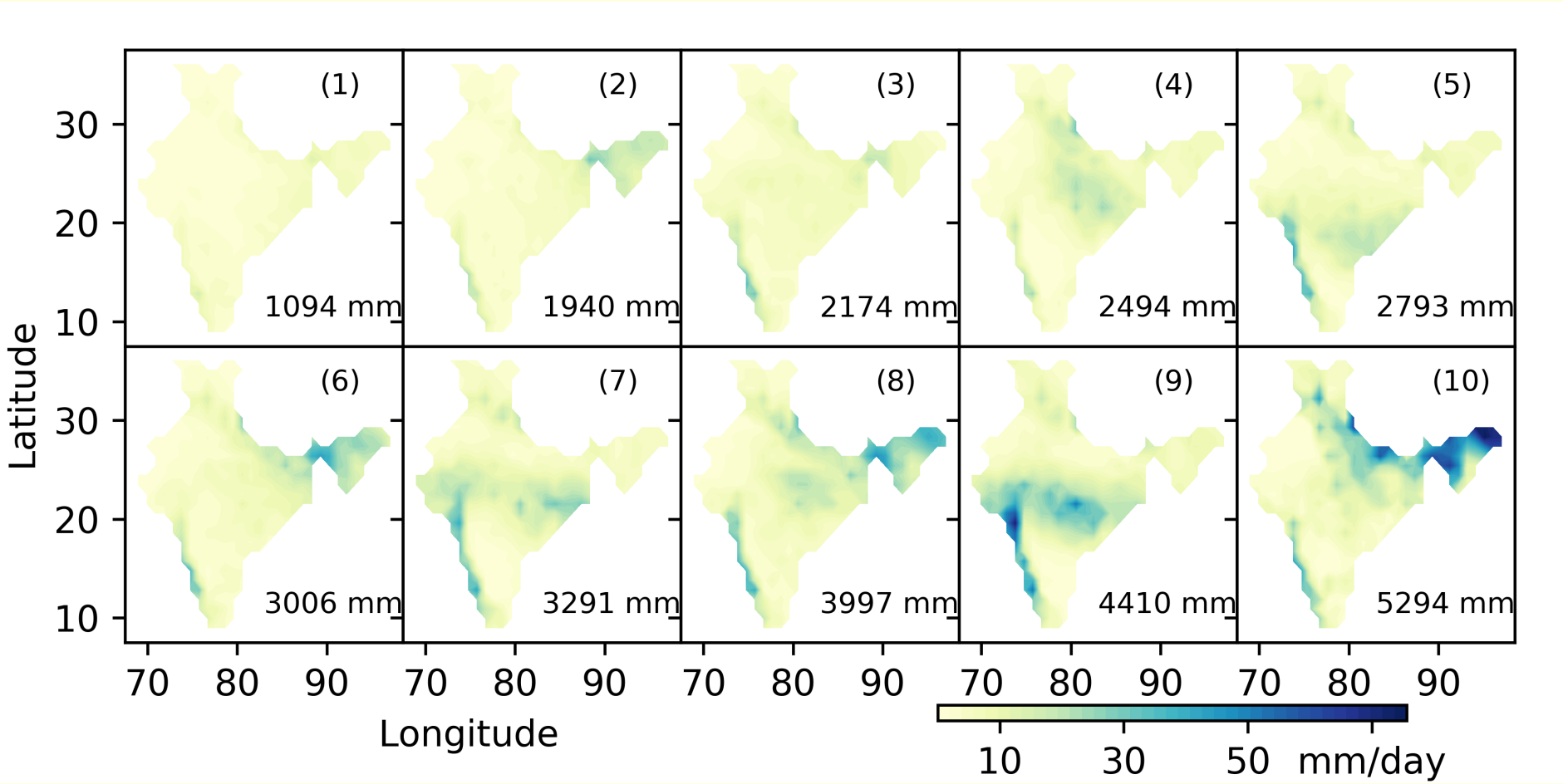
We find 10 prominent patterns

Discrete variable Z



We find 10 prominent patterns

Continuous variable X (rainfall)

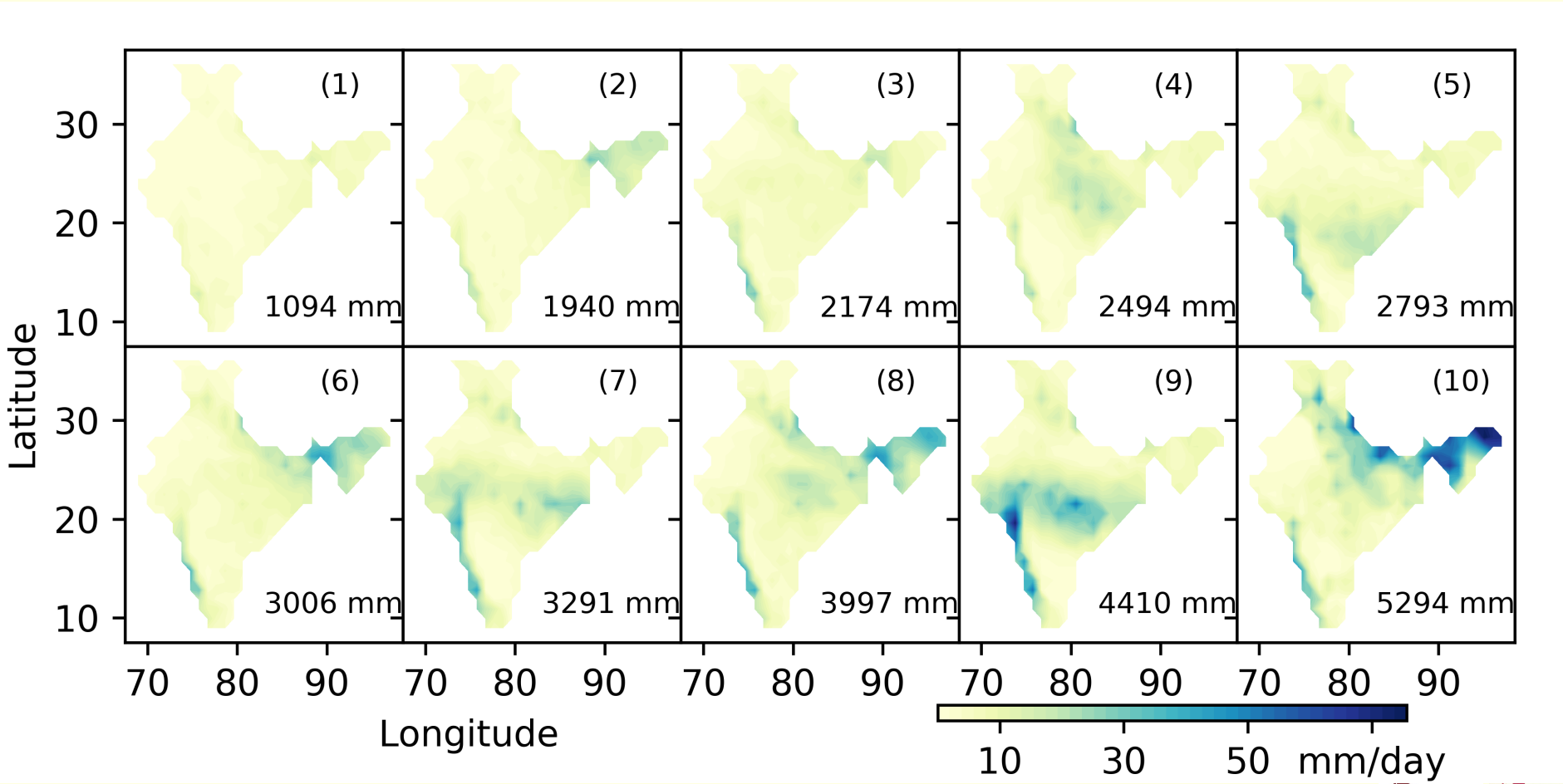


Other methods for clustering / pattern

- 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

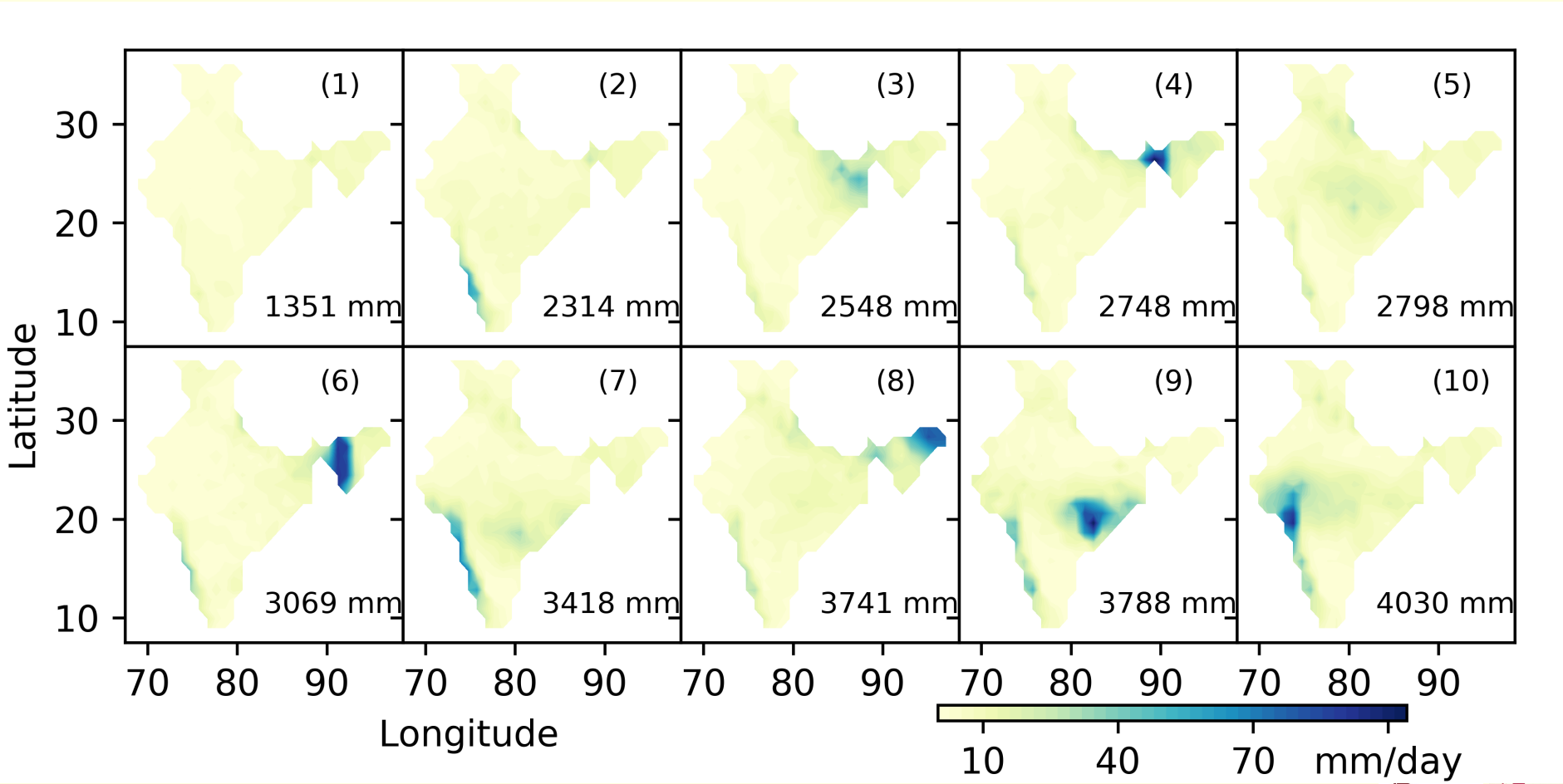
Patterns obtained by MRF are more spatially coherent and more representative

Ten patterns from MRF



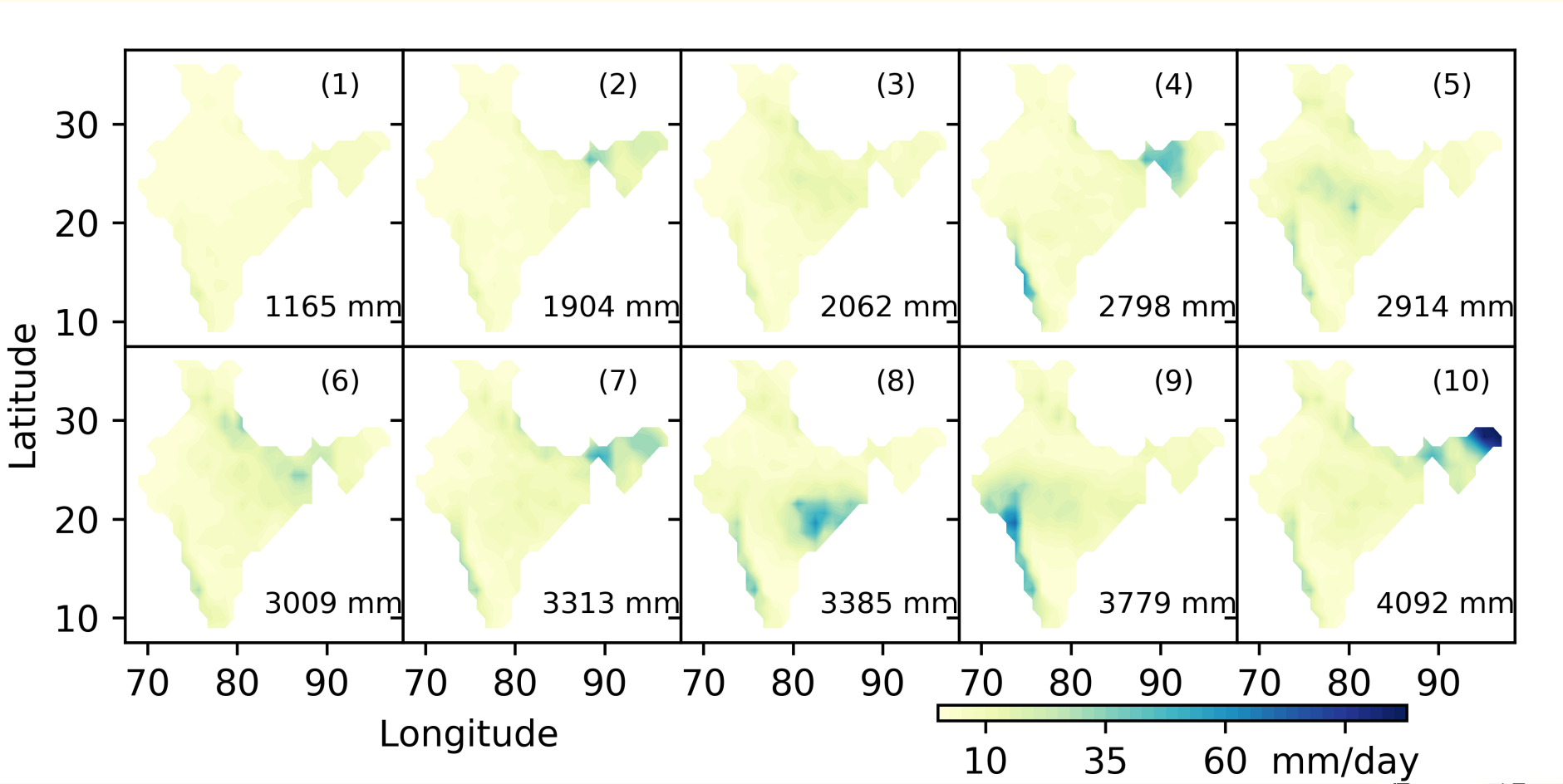
Patterns obtained by MRF are more spatially coherent and more representative

Ten patterns from K-means



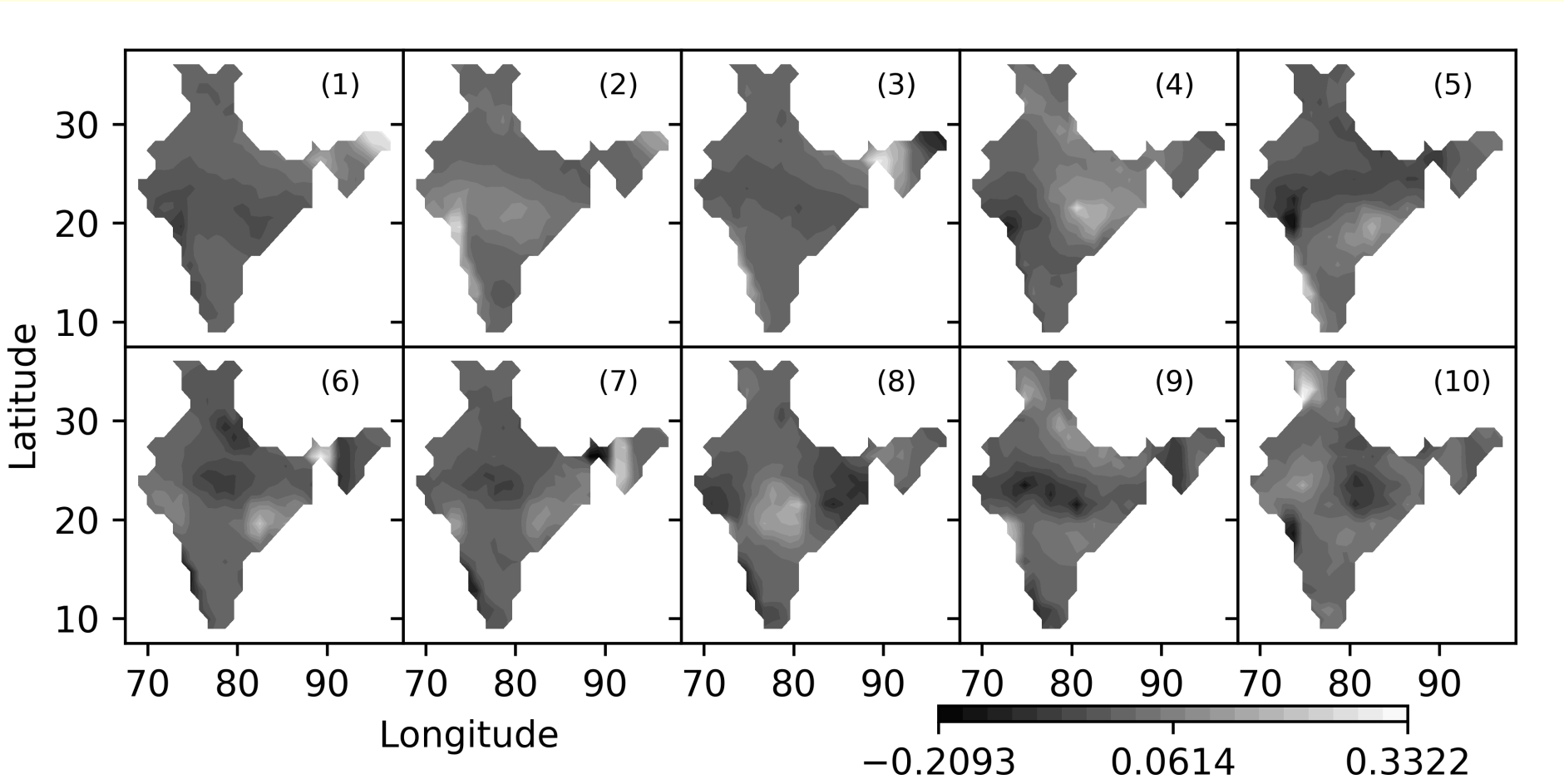
Patterns obtained by MRF are more spatially coherent and more representative

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

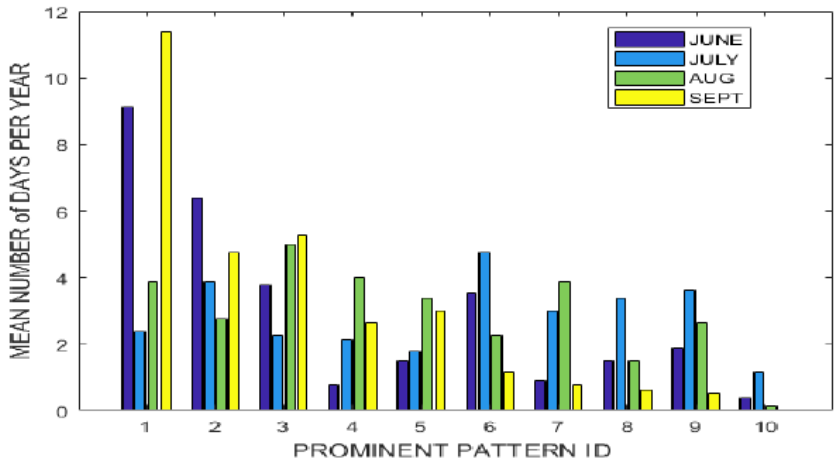
Table 5: Measures of how well the spatial patterns (CRP and CDP) computed over the period 2000-2007 can approximate the daily vectors (DRVs and DDVs) across the period 1901-2011. Three measures are considered: $\ell_2(\phi)$, Hamm(ϕ_d), and Agg(ϕ) are define in equations (14)-(16)

$spch(\phi_d)$			
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 different periods of the monsoon



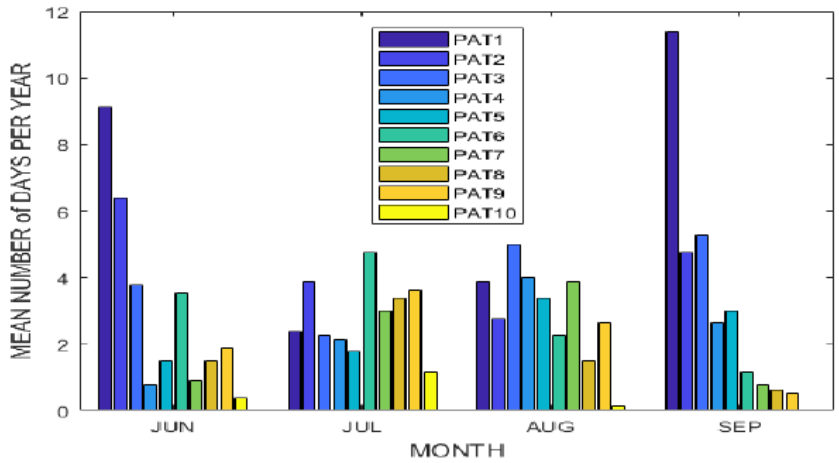


Figure 4: Left: Average number of days under each prominent pattern that belong to the 4 monsoon months (based on the period 2000-2007). Right: Average number of days in each of the 4 monsoon months that were assigned to each prominent pattern (based on the period 2000-2007).

Dynamics of these patterns

We also consider a Markov chain of these patterns: if day N is in pattern U (y-axis), what is the probability that day $N+1$ is in pattern U' (x-axis)?

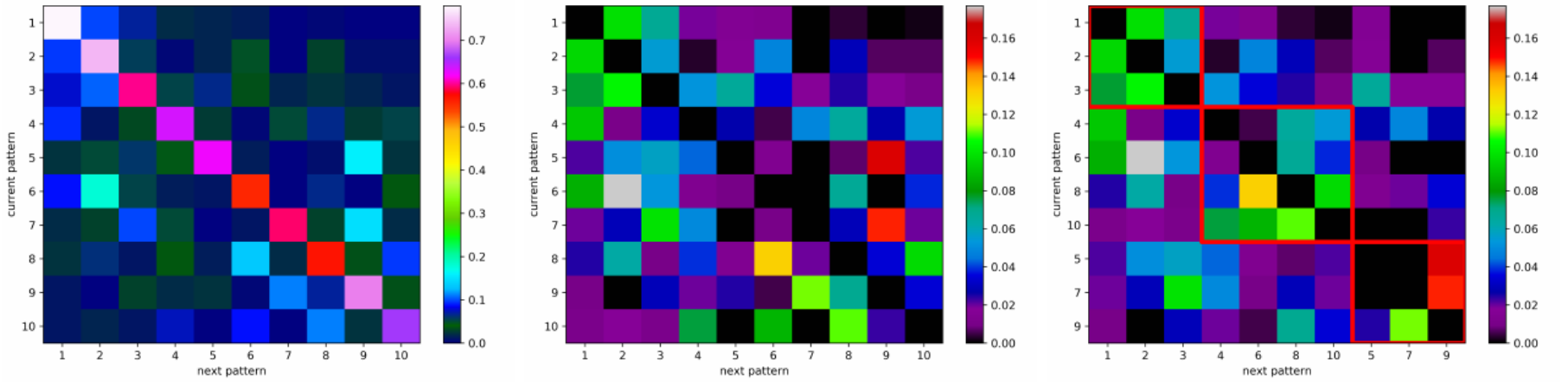


Figure 7: Transition matrix of the patterns between consecutive days. In the left matrix the diagonal elements dominate, indicating strong tendencies of self-transition in each state. In the middle matrix, the diagonal elements have been set to 0 to highlight the transitions other than same-state transitions. 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.

Dynamics of these patterns

Some transitions appear very frequently

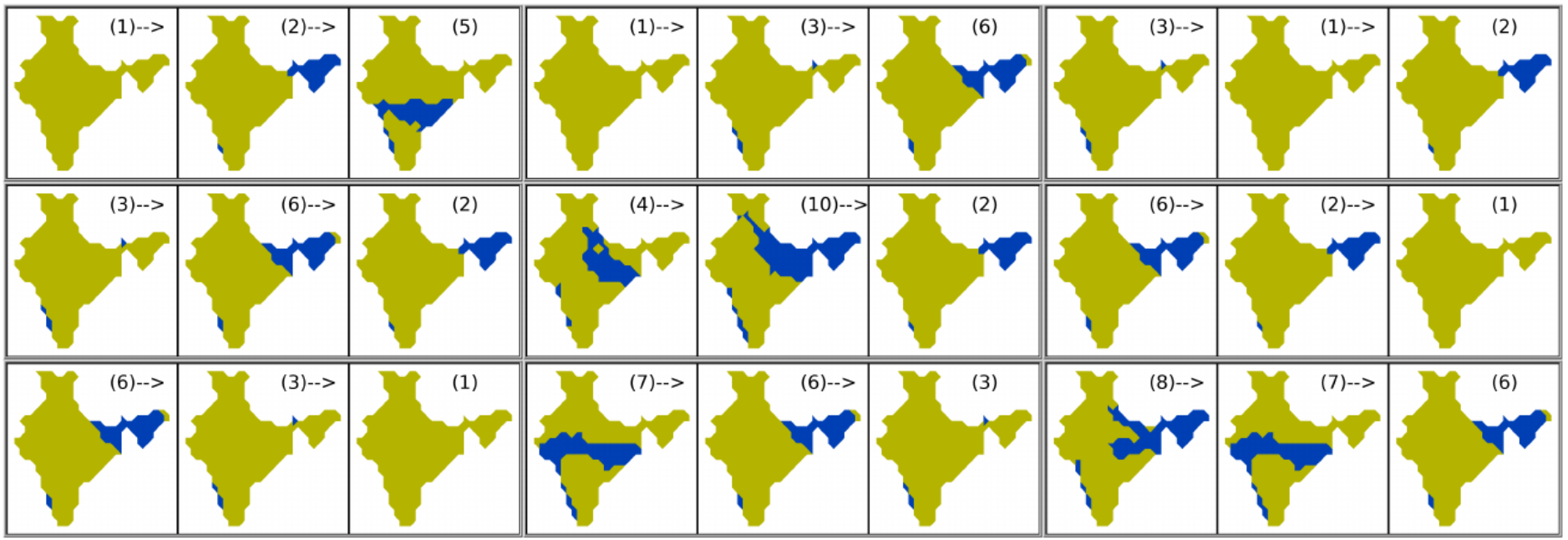


Figure 8: Frequent 3-step transition patterns among the prominent Canonical Discrete Patterns identified by the proposed model and displayed in Figure 2

Dynamics of these patterns

Summary: we construct a stochastic dynamics over the 10 patterns presented earlier, with the following properties:

Different patterns are associated to different periods of the monsoon

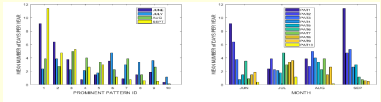


Figure 4: Left: Average number of days under each prominent pattern that belong to the 4 monsoon months (based on the period 2000-2007). Right: Average number of days in each of the 4 monsoon months that were assigned to each prominent pattern (based on the period 2000-2007).

The transition matrix is dominated by the diagonal

$\Longrightarrow$ temporal coherence / continuity

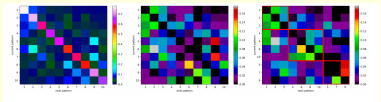


Figure 7: Transition matrix of the patterns between consecutive days. In the left matrix the diagonal elements dominate, indicating strong tendencies of self-transition in each state. In the middle matrix, the diagonal elements have been set to 0 to highlight the transitions other than same-state transitions. 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.

Figure 8: Frequent 3-step transition patterns among the prominent Canonical Discrete Patterns identified by the proposed model and displayed in Figure 2

Outline

Avenues for further exploration

- Multi-variable studies: some promising results already; extension to include vertical velocities, etc.
  - For example, to model OLR and rainfall: discrete variable Z now takes four values: low/high rainfall and low/high OLR. (Recall, there is no predetermined threshold but rather probabilistically estimated by the model.)
  - One difficulty: need to define edge potentials connecting the OLR nodes to rainfall nodes: currently there are no explicit links
- 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 improve the global circulation models