A discrete view of the Indian monsoon to identify spatial patterns of rainfall

Adway Mitra\textsuperscript{a}, Amit Apte\textsuperscript{b,\,*}, Rama Govindarajan\textsuperscript{b}, Vishal Vasan\textsuperscript{b} and Sreekar Vadlamani\textsuperscript{c,d}

\textsuperscript{a}School of Electrical Sciences, Indian Institute of Technology, Bhubaneswar, Odisha, India, \textsuperscript{b}International Centre for Theoretical Sciences (ICTS), Bangalore, Karnataka, India, \textsuperscript{c}Centre for Applicable Mathematics, Tata Institute of Fundamental Research (TIFR-CAM), Bangalore, Karnataka, India and \textsuperscript{d}Department of Statistics, Lund University, Lund, Sweden.

\textsuperscript{*}Correspondence Amit Apte, International Center for Theoretical Sciences (ICTS), Bangalore, Karnataka, India; E-mail: apte@icts.res.in

1. Introduction

The Indian summer monsoon is a multiscale, multiphysics event that has a profound impact on the food security of a billion people (Gadgil and Gadgil, 2006). Predictions for every year’s rainfall from present-day simulations are often wide off the mark (Gadgil \textit{et al.}, 2005), which has serious consequences for the economy of the country. An understanding of the basic dynamical system of the monsoon, including the role of the seasonal migration of the ITCZ and the land-ocean temperature contrast, is an important open problem in climate sciences (Sikka and Gadgil, 1980). The intra- and inter-annual variability (the temporal variation) and the spatial heterogeneity of the monsoon rainfall are quite substantial as are the deviations of local rainfall from the long-term average (Gadgil, 2003). The main aim of the work, presented here...
and in the companion article (Mitra et al., 2018), is to build a data-driven, discrete model of the daily rainfall data at 357 locations over the Indian region, with special emphasis on discovering spatially and temporally small-scale or localized properties and their relation to large-scale or annual, all-India patterns. This is achieved by constructing a Markov random field (MRF) model consisting of two main types of variables (to be described precisely later in section 2.1):

1. ‘Hidden’ discrete states (random variables Z) at each spatio-temporal location for which the rainfall data are available. The variable Z at each spatio-temporal location, and on each day, takes on one of two values, corresponding to high and low rainfall.
2. Discrete, ‘clustering’ variables U for each day and V for each location that, respectively, encode the spatial and temporal patterns in the discrete variables Z and in the data itself. The variable U for each day takes integer values and essentially indicates the spatial pattern assigned to that day (and similarly, the integer value of V indicates the temporal pattern assigned to that location).

The model we develop is probabilistic, in the sense that the main object of interest is the conditional probability distribution of the random variables Z, U, V conditioned on the rainfall data. We use Gibbs sampling algorithm (Neal, 1993) to sample this probability distribution, and the mode of this distribution is used to study the spatio-temporal patterns for the monsoon rainfall. These patterns are certainly hidden in the data, but are not easy to glean from the data directly. The probabilistic model in terms of the discrete random variables Z, U, V helps us discover these patterns. This article presents a discussion of the spatial patterns illustrating the utility of the methodology developed here. Further analysis concerning the temporal evolution of the patterns identified in this article forms one of the central aspects of the companion article (Mitra et al., 2018). In particular, we study probably transitions of one spatial pattern to another within a monsoon season. We also analyse the spatial distribution of the temporal patterns identified in this article.

The discussion in this article is organized as follows. The rest of the introduction summarizes the salient features of the model we study, followed by past data-based approaches to study patterns in spatio-temporal data in earth sciences, ending with a discussion of our main results related to spatial patterns. Then, a complete mathematical description of the MRF model is presented in detail in the following section 2. An evaluation of the spatial patterns obtained from this model compared with the patterns obtained from two other commonly used methods, K-means (Hartigan and Wong, 1979) and spectral clustering (Ng et al., 2002) is in sections 3.1–3.2, while a graphical representation of these patterns is given in section 3.3. We end with conclusions in section 4.

1.1. Main features of the present model

We construct an MRF model with four types of variables (Z, U, V, X), the first three of which are described earlier, and the fourth one X is a continuous, real-valued random variable denoting the rainfall at each spatio-temporal position. The MRF model is shown schematically in Figure 1 and described in detail in section 2.2. A novel feature of our model is the introduction of the random variables U, V that denote membership of a specific day or location to a cluster. More explicitly, recall that U takes integer values, say 1, 2, . . ., L (e.g. L = 10 later in the article). Then U(t) = k means that the day number t is assigned the kth spatial pattern out of a total of L spatial patterns. If t1, t2, . . ., td a total of d number of days which have U(ti) = k, then all these d days essentially form a cluster that have the same spatial pattern of rainfall. Similarly, all locations s1, s2, . . ., sl that are assigned the same temporal pattern V(sj) = n form a cluster. But note that the assignment of the days and locations to clusters is not deterministic, but probabilistic.

Thus, in summary, this model has the following two main characteristics:

1. The MRF for (Z, U, V, X) contains within itself an MRF for the Z variables alone. This Z-MRF has edges connecting each location to its geographical neighbours and each day to adjacent days. These edges aid in obtaining a spatio-temporally coherent picture of the rainfall.
2. The full MRF contains edges between the rainfall data X and the clustering variables U as well as edges between the discrete rainfall states Z and the clustering variables. These edges lead the model towards prominent clusters (precisely defined in section 3.2).

The edge potentials for each of these edges as well as priors on the U, V nodes are defined to take into account the qualitative features we expect for the relation between these variables, as explained in detail in sections 2.2.1–2.2.4.

Once all the edge potentials and priors on some of the variables are defined (section 2.2), the joint distribution for all the variables p(Z, U, V, X) is just a product of all the edge potentials and priors. The main inference step is
the sampling of the conditional distribution \( p(Z, U, V|X) \), conditioned on the rainfall data \( x(s, t) \), which in our case is daily rainfall for 357 locations for 8 years (2000–07). The mode of this conditional distribution is then used as an estimate of hidden state variables \( Z,U,V \). Each of the clustering variables is naturally related to a discrete as well as continuous rainfall pattern, as defined in equations (1)–(2) and summarized in Table 1.

In section 3, we discuss the properties of the mode of this conditional distribution, including the spatial rainfall patterns of this mode. Note that the conditioning on the observed data leads the model from generic coherent clusters that are present in the prior itself to the coherent clusters that are specific to the rainfall data used for conditioning. We also use these same clusters and test them against a much larger data during 1901–2007 and find that they are robust in a sense described in detail in section 3.2.

Thus, in a nutshell, the MRF model edges between \( Z \) variables lead to spatio-temporally coherent patterns, and the edges to the clustering variables lead to robust clusters that are directly informed by data because of conditioning on the observed rainfall.

1.2. Relation to past work
Identification of patterns and clusters is of course a frequently studied problem in the context of climate sciences in general, and for rainfall in particular. Specifically, in the context of the Indian monsoon, the problem of understanding the so-called active and break spells (contiguous days of above or below average rainfall) both at the all-India scale and at local scales is an important question. Break spells have been related by Ramamurthy (1969) to the disappearance of low-pressure zones and eastward winds across India, in the first formal study on break spells. In contrast, Goswami and Ajaya Mohan (2001) define these phases based on the strength of winds over the Bay of Bengal, while Krishnan et al. (2000) relate these spells to the disappearance of cloud cover over north-west and central India. Active and break spells are defined by Gadgil and Joseph (2003), Rajeevan et al. (2006) and Rajeevan et al. (2010) directly by the quantity of rainfall over the Monsoon Zone. Based on the argument that ‘intraseasonal variation’ of rainfall...
is coherent over this zone (the monsoon zone) and the average rainfall over this zone is indeed representative of the rainfall within subregions of the zone’, Rajeevan et al. (2006) consider the spatial mean rainfall of the entire zone, whereas Gadgil and Joseph (2003) divided the zone into eastern and western parts. The mean rainfall for each day is compared against the climatological mean for that date, and accordingly each day is marked as ‘active’ or ‘break’, and three or more consecutive days marked this way are identified as spells. A similar analysis was done in Krishnamurthy and Shukla (2007), though they use the all-India spatial mean instead of that over the monsoon zone. Smaller subdivisions of India are studied by Singh and Ranade (2010) to identify and describe regional dry and wet spells. Recently, Kulkarni et al. (n.d.) have defined active and break spells with respect to both rainfall and wind over south-western part of peninsula.

The problem of understanding contiguous spatial patterns of rainfall has received much less attention in the context of Indian summer monsoon rainfall. In related contexts, researchers have attempted to quantify spatial coherence in many ways, for example, for African (Moron et al., 2007), Indonesian (Haylock and McBride, 2001) and, only very recently, Indian (Moron et al., 2017) rainfall. Pai and Rajeevan (2007) use principal component analysis to select climatic variables as predictors for monsoon onset date, at least 15 days in advance of the actual onset, while Saha et al. (2016) use neural networks to find predictors for annual Indian monsoon rainfall. A commonly used method is based on the empirical orthogonal functions (EOF) of the daily rainfall vectors, as done, for example, in Gadgil and Narayana Iyengar (1980) and Suhas et al. (2013) for other rainfall data sets. Note that, in contrast to the works mentioned earlier in this paragraph, the method proposed in this article is specifically targeted to clustering by assigning each day to a spatial pattern and each location to a temporal pattern.

Recent developments in machine learning and data science, along with the availability of high-resolution and accurate climatic data, have resulted in machine learning methods being used to answer questions in climate science in a data-driven way. Modelling extreme events of precipitation has been attempted using Bayesian methods, such as Sang and Gelfand (2009) and Ghosh and Mallick (2011). MRFs (Kindermann and Snell, 1980) are a natural way of handling spatio-temporal data, and they have been used for analysis of ocean surface temperature by Lavine and Lozier (1999) and detection of large-scale droughts by Fu et al. (2012). A recent work that is quite relevant to the current work is Moron et al. (2015) where six weather types, each specified by a spatial pattern of daily low-altitude horizontal winds, are identified using K-means clustering technique over the Pacific.

In the case of the Indian monsoon, not much analysis has been done using machine learning methods, and one of the main aims of the present article is to fill this gap by providing a robust way of identifying important spatial patterns of the Indian summer monsoon rainfall, using an MRF model. Our approach can be compared with a series of works (Greene et al., 2008, 2011; Moron et al., 2012; Holsclaw et al., 2017) where hidden Markov model has been used to identify states of rainfall over India. The state variables in these works can be considered analogous to our U variables, introduced later in section 2.1. These articles study mean spatial distribution of rainfall for each state, which roughly corresponds to what we call canonical rainfall pattern introduced in section 2, equation (1). An important difference between the proposed model and the above-mentioned HMM-based models is that the latter ones are parametric, i.e. number of possible states is fixed by the user or chosen based on goodness of model fit, unlike the proposed model which chooses the appropriate number of patterns as part of the inference process. Other novel features of our approach are that our model reinforces spatial coherence in the patterns and also allows us to study the temporal evolution of these patterns.

1.3. A summary of our approach and the main results

We use daily rainfall data of the monsoon seasons (1 June–30 September) of 8 years (2000–07) over 357 locations across India to obtain a few prominent spatial patterns. The spatial distribution of each day’s rainfall is described by one pattern. These patterns need to be chosen so as to simultaneously (i) minimize the difference between each day’s rainfall distribution (in a discrete representation, as described later) and that of its representative pattern and (ii) reduce the total number of patterns needed to describe the rainfall distribution on all the days. In other words, the aim is to obtain a small number of rainfall patterns that are able to represent the spatial rainfall distribution on all days. In a sense, this is analogous to using leading singular vectors to create a concise representation of a large data set.

Daily rainfall at a given location is of course a manifestation of an extremely complicated dynamical process involving many physical processes affected by global atmospheric and ocean conditions. This gives rise to the expectation that rainfall over short distances and times will be correlated to each other. To incorporate this in the simplest
In our model, we use a discrete representation of each day’s rainfall at each location. In the MRF model, we impose edge potentials between neighbouring spatial locations and times, which has the effect of enhancing spatio-temporal coherence of the discrete variables.

We find that our approach discovers 10 spatial patterns, and these patterns are robust over a range of user-chosen parameters. More interestingly, the patterns we obtain from the 8-year data (2000–07) are representative (to within errors as described later) of over 95% of all days in the monsoon seasons of 107 years, from 1901 to 2007. The spatial patterns discovered by our model are more coherent, comprehensive and physically realistic compared with other approaches such as K-means, spectral clustering and EOF.

To demonstrate that our MRF approach is a preferable option to identify spatial patterns in the Indian monsoon rainfall, we obtain for comparison spatial patterns, for the same 8-year rainfall data, by several other approaches, such as by EOF (Sikka and Gadgil, 1980), K-means (Hartigan and Wong, 1979) and spectral means (Ng et al., 2002). The EOF are not a direct way of visualizing daily rainfall, since the pattern on a given day is obtained as a linear combination of several modes.

When compared with K-means and spectral clustering, the MRF model gives the least number of prominent patterns, which are able to approximate the rainfall vectors for the largest number of days in the data set and are thus a much better sparse representation of the data compared with the other methods. The MRF patterns are also the most coherent in real space, in the sense that adjacent positions are more likely to have similar values of the discrete rainfall variable. Assignment of each day to one of these patterns by the MRF creates a clustering of the days (indicated by variable U). These clusters also cover the largest fraction of the total number of days, compared with clusters of K-means and spectral approaches. The MRF clusters also show the smallest mean Hamming distance of the daily discrete rainfall patterns from that of the cluster each day belongs to, but all the methods show approximately same mean $\ell_2$ distance. Interestingly, when the discrete patterns identified from 2000 to 2007 data are used for clustering of the data from 1901 to 2007, the MRF clusters show the smallest mean Hamming and $\ell_2$ distances. Since the hamming distance is a natural metric to evaluate discrete patterns, this indicates that our clusters are coherent in the data space.

The patterns we identify, and the sequence of their occurrence, provide an ideal platform to understand many features of the monsoon, and we make a beginning in the companion article (Mitra et al., 2018). In that article, we collate the results of this article, i.e. the spatial and temporal patterns, and thoroughly analyse various aspects of these patterns.

2. Methodology

In this section, we describe the mathematical model based on MRF and our main motivation for this model. We also present a description of other methods with which we compare the results from our MRF study. Note that this whole section is devoted to introduce the notation and the model before we come to describing the results in section 3.

Consider $S$ locations, which correspond to grid-cells on a rectangular grid system, ordered sequentially first according to longitude and then according to latitude. The indexing scheme has no bearing on the analysis that follows. For each location $s$ with coordinates $(\ell_1, \ell_2)$ we define $\Omega(s)$ as the neighbourhood of this location with coordinates $(\ell_1 + a, \ell_2 + b)$ where $a \in [-1,0,1]$ and $b \in [-1,0,1]$ (except for the points on the boundary for which we have to appropriately drop one of $\pm 1$). Consider a sequence of $T$ time points, each of which corresponds to a day. These days can belong to different years, and the year of day $t$ is indicated by $YY(t)$.

The observed rainfall at location $s$ on day $t$, denoted by $x(s,t)$, is arranged as a matrix of dimension $S \times T$. Corresponding to each day $t$, we have an $S$-dimensional ‘daily rainfall vector’ (DRV) $x(t)$ which is a column of that matrix. Similarly, each location $s$ is associated with a ‘rainfall time series’ (RTS) $x(s)$, which is a row of the matrix.

We want to find a set of ‘canonical rainfall patterns’ (CRPs) $\{\phi_k\}$ each of which is an $S$-dimensional vector indicating a specific spatial pattern of the rainfall, with the aim of expressing each DRV $x(t)$ in terms of the CRPs. For example, in our model, if $U(t) = u$, then the day $t$ is assigned to the pattern $\phi_u$, whereas in methods such as EOF (described in detail later), $x(t)$ is written as a linear combination of $\{\phi_k\}$. We then wish to identify transition patterns among these CRPs, with the aim of providing a concise view of the progress of monsoon in terms of the temporal evolution of the CRPs. We also want to cluster the days according to the spatial patterns of rainfall and aggregate rainfall, such that some of the clusters belong to ‘active days’, some to ‘break days’ and so on. To achieve that task, we want to find a set of ‘canonical time series’ (CTS) $\{\theta_l\}$ each of which is a $T$-dimensional vector indicating a specific temporal pattern of the rainfall, with the aim of expressing each RTS $x(s)$ in terms of the CTSs, in analogy of relating DRVs and CRPs.
We now present a novel approach to the tasks described earlier. At the heart of this approach is a discrete representation that is easy to interpret and visualize and allows us to study properties of monsoon rainfall at local and regional scales.

2.1. Discrete representation: notation and definition

Consider a binary latent (random) variable \( Z(s, t) \in \{1, 2\} \), which encodes the rainfall amount \( x(s, t) \) at location \( s \) on day \( t \). Its two states \( \{1, 2\} \) parameterize two types of distributions for rainfall: \( Z(s, t) = 1 \) corresponds to ‘high’ rainfall, and \( Z(s, t) = 2 \) corresponds to ‘low’ rainfall at the location \( s \) on day \( t \). Note that we do not put any \textit{a priori} threshold to pre-define the high and low rainfall amounts, but the thresholds are estimated \textit{a posteriori}. Effectively, the daily rainfall \( X \) at each location is modelled using a two-component mixture distribution, where one component has high expected value and hence is also an \textit{indicator} for location \( s \) indicates the membership of location \( s \) when the day \( t \) corresponds to ‘high’ rainfall, \( V(s) = 1 \). (, , )

The main objective of introducing these new random variables is to define the CRPs and the CTSs, and their associated discretized versions CDPs and CDSs that were mentioned earlier. This is done as follows: for each temporal index \( u \) (i.e. for each temporal cluster indexed by the integer \( u \)), the associated temporal patterns CTS and CDS are defined by

\[
\phi_u = \text{mean}(X(t) : U(t) = u), \\
\phi'_u = \text{mod}(Z(t) : U(t) = u),
\]

(1)

where the mean/mode is taken over the days that belong to the cluster \( u \). Recall that each \( X(t) \) is an \( S \)-dimensional vector, and hence each \( \phi_u \) is also an \( S \)-dimensional vector. Similarly, for each spatial index \( v \) (i.e. for each spatial cluster indexed by the integer \( v \)), the associated spatial patterns CRP and CDP are defined by

\[
\theta_v = \text{mean}(X(s) : V(s) = v), \\
\theta'_v = \text{mod}(Z(s) : V(s) = v),
\]

(2)

where the mean/mode is taken over the locations that belong to the cluster \( v \). Again, \( X(s) \) and hence \( \theta_v \) are \( T \)-dimensional vectors.

2.2. MRF model

So far, we have introduced four sets of random variables \( (X, Z, U, V) \): the rainfall \( X(s, t) \) and a discrete rainfall state \( Z(s, t) \) at each location \( s \) on day \( t \); the cluster membership indicated by \( U(t) \) for day \( t \) and \( V(s) \) for location \( s \). The observational rainfall data \( x(s, t) \) is a realization of the random variables \( X(s, t) \). We aim to make inferences about the discrete variables \( Z, U, V \) conditional on the observed data, which can be written as follows:

\[
p(Z, U, V|X) = \frac{p(Z, U, V, X)}{p(X)} \propto p(Z, U, V, X).
\]

(3)
To study this conditional distribution, we will use Gibbs sampling method, and hence we will not need the normalization constant \( p(X) \) explicitly. Thus the main task now is to describe the joint distribution \( p(Z, U, V, X) \) whose definition is in terms of an MRF with \( Z(s, t), U(t), V(s), X(s, t) \) as its nodes. MRF is a graphical model (see Kindermann and Snell, 1980) where each random variable is represented by a node, and some pairs of nodes are connected by edges, each of which is associated with an edge potential, and the joint distribution of all the random variables is the product of these edge potentials. In our specific MRF for the \( (Z, U, V, X) \) variables, this joint distribution is summarized in the equation below, and the MRF itself is shown schematically in Figure 1. Thus, the joint density \( p(Z, U, V, X) \) is summarized in the following equation:

\[
p(Z, U, V, X) \propto p_U(U) \times p_V(V) \times \prod_{t=1}^{S,T} \prod_{s,t} \psi_T(Z(s, t), Z(s', t)) \\
\times \prod_{s,t} \prod_{s', t' \in \theta(s,t)} \psi_G(Z(s, t), Z(s', t)) \\
\times \prod_{s,t} \psi_{ST}(Z(s, t), V(s)) \times \prod_{s,t} \psi_{SS}(Z(s, t), U(t)) \\
\times \prod_{s,t} \psi_{SU}(Z(s, t), X(s, t)) \times \prod_{t} \psi_{DU}(U(t), Y(t))
\]

In equation (4), the first two lines contain the prior distributions on the \((Z, U, V)\) nodes (the prior on \( Z \) itself being described in terms of the edge potentials for the MRF on \( Z \) nodes alone), defined in equations (5)–(7). The third line indicates the edge potentials on the \( Z - U \) and the \( Z - V \) edges, as defined later in equation (8). The last line indicates the edge potentials on the \( X - Z \) and the \( X - U \) edges, as defined later in equations (9)–(10). Recall that \( Y \) is just a shorthand for the sum of \( X \) along the spatial locations and not a new random variable. We now define in detail each of the terms in equation (4).

### 2.2.1. Prior for the clustering variables \( U, V \)

As stated earlier, the number of clusters is equal to the number of unique values taken by the \( U \) and \( V \) variables, but these are not decided beforehand. Instead we invoke Chinese restaurant process priors on both \( U \) and \( V \). This is a popular and simple approach to Bayesian nonparametric clustering expressed as a sequential process (Pitman, 1995). The day \( t \) can join a cluster from \( \{U(1), \ldots, U(t-1)\} \) with probability proportional to the number of days assigned to that cluster so far, or form a new cluster with probability proportional to \( \gamma \), which is a tuning parameter to be chosen by us. We modify the method slightly, so that the probability of using an existing cluster \( k \) is the number of days assigned to it, times the number of years where at least 1 day was assigned to it. Denote by \( n_{tu} \) the number of days assigned to cluster \( u \) so far, and \( m_{tu} \) the number of years where at least 1 day was assigned to cluster \( u \). Then, the prior distribution of \( U \) is given by the following conditional distribution for each \( U(t) \) variable:

\[
P(U(t) = u | U(1), \ldots, U(t-1)) \propto \begin{cases} 
n_{tu}m(t,u) & \text{if } u \in \{U(1), \ldots, U(t-1)\}, \\
\gamma & \text{for } u = \max(U(1), \ldots, U(t-1)) + 1,
\end{cases}
\]

\[
P(U(1) = 1) = 1.
\]

By modification of Chinese restaurant process, we ensure that the prominent clusters contain days from different years, i.e. are general across the years rather than forming year-specific clusters. This distribution has the ‘rich getting richer’ property, i.e. each variable is likely to be assigned to a cluster that is already large. This helps in the identification of a few prominent patterns. We note that this prior distribution implied by the Chinese restaurant process does not take into account any information about the actual rainfall on any given day (i.e. the random variable \( X \)).
For the spatial cluster variables $V$, we again make use of a similar prior distribution based on Chinese restaurant process, where each location $s$ can join a cluster from $\{V(1), \ldots, V(s-1)\}$ with probability proportional to the number of locations assigned to that cluster so far (denoted by $n(s, v) = \sum_{v'} : s_v < s, V(s_v) = v'$), or form a new cluster with probability proportional to $\lambda$. Once again, this helps in identification of a few prominent series.

$$
P(V(s) = v|V(1), \ldots, V(s-1)) \propto \begin{cases} 
n(s, v) & \text{if } v \in \{V(1), \ldots, V(s-1)\}, \\
\lambda & \text{for } v = \max[V(1), \ldots, V(s-1)] + 1.
\end{cases}$$

Let us denote the densities for these prior distributions of $U$ and $V$ as $p_u(U)$ and $p_v(V)$.

### 2.2.2. Prior for the discrete rainfall variables $Z(s,t)$

We put an MRF prior on the $Z$ variables. In this case, we consider each $Z(s, t)$ variable as a node. For every location $s$, we consider spatial edges to the nodes representing its neighbouring locations in $\Omega(s)$, i.e. $Z(s, t)$ and $Z(s', t)$ are connected by an edge for each $s' \in \Omega(s)$. Similarly we consider temporal edges connecting successive days, i.e. $Z(s, t)$ and $Z(s, t+1)$ are connected by an edge for each location $s$ and each day $t$. The edge potential functions on all these edges are described below.

We have already discussed that most climatological phenomena, and especially rainfall, are spatio-temporally coherent. This is particularly true of the discrete (binary) state variable, since it represents the ‘state of the climate’ which is a large-scale concept. Hence it is expected to exhibit significantly more spatial and temporal coherence than the actually measured rainfall itself. So, we define the potential functions of temporal and spatial edges between the state variables to promote spatial and temporal coherence. Specifically, we set each function to take high values if the $Z$ variables connected by the edge are equal and low values if they are different, as given in the equation below.

$$
\begin{align*}
\psi_f(Z(s, t), Z(s, t+1)) &= f 1_{[Z(s, t) = Z(s, t+1)]}, \\
\psi_d(Z(s, t), Z(s', t)) &= g(s, s') 1_{[Z(s, t) = Z(s', t)]} \text{ for } s' \in \Omega(s). 
\end{align*}
$$

Here $1$ is the indicator function and $s'$ is the neighbourhood of $s$. The temporal-coherence parameter $f$ is assumed constant, while the spatial coherence parameter $g$ is specific to the pair of locations $(s, s')$ because not every pair of neighbouring locations has the same degree of correlation. For example, locations on the western (windward) side of the Western Ghats mountain range receive very heavy rain during monsoon compared with those on the eastern (leeward) side, even though they are in adjacent grid points, because the mountain range is narrower than the dimensions of our grids. We set $g(s, s')$ to be equal to the correlation coefficient between the RTS at the two locations $(X(s)$ and $X(s'))$.

### 2.2.3. Edge potentials between discrete rainfall variable $z$ and clustering variables $U, V$

We also construct spatial scale edges between the discrete rainfall variables ($Z$ nodes) of each day to the daily cluster variable $U$ for that day, and temporal scale edges between the $Z$ nodes of each location to the spatial cluster variable $V$ for that location. The edge potential functions defined on these edges try to align the spatial $Z$-vector of each day (daily discretized vector DDC) to the CDP ($\phi^d$-vector) associated with that day’s spatial cluster, and the temporal $Z$-series of each location [discretized time series (DTS)] to the CDS ($\theta^d$-vector) associated with that day’s temporal cluster, as described in equations below.

$$
\begin{align*}
\psi_{3d}(Z(s, t), U(t)) &= \exp\left(\eta 1_{[Z(s, t) = \phi^d(s, U(t))]}\right), \\
\psi_{5d}(Z(s, t), V(s)) &= \exp\left(\zeta 1_{[Z(s, t) = \theta^d(V(s), t)]}\right).
\end{align*}
$$

### 2.2.4. Edge potentials between rainfall data $X$ and discrete rainfall and clustering variable $Z, U$

We have already mentioned that each $Z$ variable encodes a distribution over the observations $X$. In particular, we choose $X(s, t) \sim \Gamma(\alpha_x, \beta_x)$ when $Z(s, t) = z$. Thus, for each of the two values $z = 1, 2$, the distribution of $X$ has different
means $\alpha_1$ and $\alpha_2$ and also different variance. Thus, we naturally include these $X$ variables as nodes into our graphical model by adding data edges between each $Z(s, t)$ and $X(s, t)$ node and defining the edge potential of these edges to be the same as the above data distributions, i.e. these gamma PDFs.

$$
\psi_{DZ}(Z(s, t) = z, X(s, t)) = (X(s, t))^{\alpha_{zt} - 1} \exp(-\beta_{zt} X(s, t)).
$$

(9)

This gives a direct way for the data to inform the $Z$ variables, when we sample the conditional distribution $p(Z, U, V|X)$. Similarly, we would like to data to directly influence the spatial clustering and hence we include in our MRF the edges between the rainfall variables $X(s, t)$ for a fixed $t$ to the clustering variable $U(t)$ for the same day. To describe this edge potential, it is easiest to introduce a variable $Y(t) = \sum_{s=1}^{S} X(s, t)$. This aggregate rainfall is chosen to have a Gaussian PDF with the mean and covariance $(\mu_u, \sigma_u)$ that depends on the cluster $U(t) = u$ to which the day $t$ belongs. Thus, each daily cluster is associated with a distribution of the daily aggregate rainfall.

$$
\psi_{DU}(U(t), Y(t)) = \exp\left(-\frac{1}{2} \frac{(Y(t) - \mu_{u(t)})^2}{\sigma_{u(t)}^2}\right).
$$

(10)

This completes our description of the MRF for the $(Z, U, V, X)$ variables and their joint distribution given in equations (4)–(10). The various parameters that occur in this description are summarized in Table 2, indicating which of them are set by the user and which are inferred from the model.

### 2.3. Model inference by Gibbs sampling

Having defined the model with all the random variables and parameters, we now come to the main step: inference of the latent variables $(Z, U, V)$ and estimation of the parameters: $(\alpha, \beta, \mu)$. The other parameters $(f, g, \eta, \zeta, \gamma, \lambda, \sigma)$ are left as user-defined (see Table 2). The idea of inference is to sample the conditional distribution $p(Z, U, V|X)$ as defined in equation (4) and using the samples, find the assignment of parameters that maximizes the likelihood.

We use Gibbs sampling (Neal, 1993), which is a Markov chain Monte Carlo approach. Here, we start with an initial assignment to all the latent variables. Then, we visit each random variable one by one, and assign it a value sampled from its conditional distribution, based on all the remaining latent variables. This process is repeated to obtain samples from the posterior distributions of the latent variables. We use the mode (obtained from the samples) of this posterior distributions as the optimal estimate of the latent variables.

In our case, we also have many parameters of these distributions, which are unknown. Estimating these parameters by expectation–maximization is computationally very costly. Instead, we use maximum likelihood estimation simultaneously with Gibbs sampling, where each iteration of latent variable inference is followed by taking a maximum-likelihood estimate of all the parameters based on the current assignment of all the latent variables.

### Table 2. The various parameters and the corresponding mathematical symbols related to the proposed model used throughout this and the companion article

| Symbol | Dimension | Defined in | Description | Set by |
|--------|-----------|------------|-------------|--------|
| $\gamma$ | 1 | Equation (5) | Parameter of Chinese restaurant process on $U$ | User |
| $\lambda$ | 1 | Equation (6) | Parameter of Chinese restaurant process on $V$ | User |
| $f$ | 1 | Equation (7) | Temporal coherence parameter on $Z$ | User |
| $g$ | $S \times \Omega$ | Equation (7) | Spatial coherence parameter on $Z$ | User |
| $\eta$ | 1 | Equation (8) | Relation between $Z$ and $U$ | User |
| $\zeta$ | 1 | Equation (8) | Relation between $Z$ and $V$ | User |
| $\alpha$ | $2S$ | Equation (9) | Shape parameter of gamma distribution on $X$ | Model |
| $\beta$ | $2S$ | Equation (9) | Scale parameter of gamma distribution on $X$ | Model |
| $\mu$ | $K^a$ | Equation (10) | Mean of Gaussian distribution on $Y$ | Model |
| $\sigma$ | $L^a$ | Equation (10) | Standard deviation of Gaussian distribution on $Y$ | User |

$^a(K/L)$ is the number of patterns/clusters created by the model.
Next, we describe the sampling of each random variable: \(Z(s, t), U(t)\) and \(V(s)\). In each case, the conditioning set is all the remaining random variables. Here the Markov properties of the MRF (see Kindermann and Snell, 1980) become important: each variable is independent of all other variables conditioned on its neighbours. This makes the Gibbs sampling step vastly simpler: for sampling, each variable we can drop all the terms that do not involve its neighbouring nodes.

For sampling \(Z(s, t)\), the distribution is as follows:

\[
\mathbb{P}(Z(s, t) = z|\bar{Z}, U, V, X) \propto \prod_{t' \neq t} \psi_t(z, Z(s, t')) \times \prod_{t \in \Omega(s)} \psi_{3}(z, Z(s', t)) \times \psi_{ST}(z, V(s)) \times \psi_{SS}(z, U(t)) \times \psi_{DZ}(z, X(s, t))
\]

where \(\bar{Z}\) denotes all the \(Z\) variables except the node \(Z(s, t)\) being sampled.

For sampling \(U(t)\) and \(V(s)\), the distributions are as follows:

\[
\begin{align*}
\mathbb{P}(U(t) = u|Z, U, V, X) & \propto T'_u(U(t), u) \times \prod_{s} \psi_{SS}(Z(s, t), u) \times \psi_{DU}(U(t), Y(t)) \\
\mathbb{P}(V(s) = v|Z, U, V, X) & \propto T'_v(V(s), v) \times \prod_{s} \psi_{SY}(Z(s, t), v)
\end{align*}
\]

where \(\bar{U}, \bar{V}\) denote all the \(U, V\) variables except the node being sampled, and \(T'_u\) and \(T'_v\) are conditional distributions based on Chinese restaurant process, which take the same form as \(p_u\) and \(p_u\) from equations (5)–(6), considering value assignments to all the random variables, due to the complete exchangeability property of Chinese restaurant process, as explained in Pitman (1995).

Finally, the parameters \((\alpha, \beta, \mu)\) are estimated at each iteration using maximum likelihood, based on the current assignments of the random variables. For example, \(\mu_u\) is the sample mean of \(Y\) on those days that have been assigned to daily cluster \(u\), while estimates of the gamma parameters \(\alpha_{sk}, \beta_{sk}\) are obtained using the estimated mean and variance of \(X_s\) in those days when \(Z_{ks} = k\).

As mentioned earlier, the continuous and discrete canonical rainfall patterns \(\phi_u\) and \(\phi_d\) and the continuous and discrete canonical time series \(\theta_v\) and \(\theta_d\) (see Table 1), which are main quantities of interest, are obtained from equations (1)–(2) with the values of the discrete variables \((Z, U, V)\) obtained from the sample mode, and the values of the rainfall are of course obtained from the data.

2.4. Related methods

The tasks of identification of ‘canonical’ vectors as well as of patterns and clusters have been studied in various contexts, and many different methods have been proposed in data mining literature. We use some of these methods and compare the results obtained from these methods with those obtained from the MRF model described earlier.

2.4.1. Using the sample covariance matrix to obtain canonical vectors

A commonly used method is based on the EOF of the DRVs, as done, for example, in Gadgil and Narayana Iyengar (1980) and Suhas et al. (2013) for other rainfall data sets. This process gives us \(S\) vectors, each of dimension \(S\), which are the eigenvectors of the sample covariance matrix of the DRVs. We can denote these vectors as \(\{\phi_u^E, \phi_v^E, \ldots, \phi_{S}^E\}\), which can serve as CRPs. These eigenvectors are indexed in descending order of their associated eigenvalues.

Each DRV can be expressed as a sum of these eigenvectors and the mean, i.e. \(X(t) = \mu + \sum_{j=1}^{S} \alpha_j \phi_j^E\), where \(\alpha\) are regression coefficients and \(\mu\) is the mean of \(X(t)\). If we want to represent each DRV as a sparse combination of \(p\) or less CRPs, we need to solve a problem of sparse linear regression, popularly called LASSO. This can be achieved by solving a regression problem while putting an \(\ell_1\)-norm regularizer on the regression coefficients. This is formulated as follows:

\[
\min \alpha \left( \left\| X_t - \sum_{j=1}^{S} \alpha_j \phi_j^E \right\|^2 + \lambda \sum_{j=1}^{S} |\alpha_j| \right)
\]
The parameter $\lambda$ regulates the sparsity of the $\alpha$ vector, i.e. the number of non-zero entries. Increasing $\lambda$ increases the sparsity.

2.4.2. Identification of clusters to obtain canonical vectors

Another approach to identifying CRPs from the data would be to perform clustering of the data. In particular, we focus on two clustering algorithms: $K$-means (Hartigan and Wong, 1979) and spectral clustering (Ng et al., 2002) to partition all the $T$ DRVs $X(t)$ into $K$ clusters, and the $S$ RTSs $X(s)$ into $L$ clusters. From these clusters, we compute the cluster mean vectors to serve as CRPs and CTSs.

$K$-means directly uses the DRVs. Spectral clustering is done in two ways: once using the exponentials of negative Euclidean distances between DRVs (denoted by Spect1 later) and once using Hamming similarity between thresholded DDVs (denoted by Spect2). Note that the number of clusters to be formed needs to be given as an input to both these methods.

3. Results

Having defined our model, we now describe our experiments and present our results. The data set used for this work was published by Indian Institute of Tropical Meteorology. It provides daily rainfall data for the period 1901–2011 at $1^\circ$ spatial resolution (Rajeevan et al., 2006) all over India. The data set is available on request from http://www.tropmet.res.in/Data. We use daily rainfall data for 8 years, 2000–07, during the 4 months (June–September) (122 days each year), over 357 locations across India to identify the CRPs and CDPs. So in our experiments, $S = 357$ and $T = 122 \times 8 = 976$. We analyse the daily clusters $U$ and their associated canonical discrete patterns. We also approximate the DRV for an extended period of 111 years (1901–2011) during the four monsoon months over the same 357 locations using the prominent patterns we identify in this article, and the analysis of the spatial clusters $V$ is presented in a companion article (Mitra et al., 2018).

3.1. Evaluation of daily clusters and spatial patterns

As already discussed, the number of clusters is not fixed by the user, but learned from the data. However, the user has a control over this number through the parameter $\eta$ (equation (8)). A small value of $\eta$ indicates a large number of clusters, though only a few of them will be prominent (with a significant number of days assigned, and spanning across multiple years). A larger value of $\eta$ will create fewer clusters, and for sufficiently large $\eta$ all days will collapse into a single cluster. In our experiments, we vary this parameter from 5 to 10. The number of clusters obtained for each of these values of $\eta$ is shown in Table 3.

Even for small values of $\eta$, not all the clusters are significant enough, in the sense that the number of days belonging to a cluster may be too small. Hence, we define prominent clusters to be those having members from at least 5 of the 8 years. We see that even with different $\eta$ and hence different number of total clusters, the number of prominent
clusters is approximately constant at around 10. For comparison of these results with those from K-means (Hartigan and Wong, 1979) and spectral clustering (Ng et al., 2002) (which require specification of number of clusters), we specify the same number of clusters as obtained for that specific value of \( \eta \) and then find the number of prominent clusters using the same definition above. The number of prominent clusters for each of these methods (our model, K-means, and two spectral clustering algorithms denoted by SP1 and SP2, as mentioned in section 2.4) is reported under the columns titled ‘#PC’ in Table 3.

The number of days (out of a total of \( T = 976 \)) covered by the prominent clusters identified by each method indicates the significance of these clusters. This number is shown under the columns titled ‘PC coverage’ in Table 3.

An important criteria to evaluate clustering is the intra-cluster homogeneity. In this case, it is expected that each cluster should be uniform with respect to the daily aggregate rainfall denoted by \( Y(t) \). A cluster’s uniformity with respect to \( Y \) can be measured by standard deviation of the \( Y \)-values assigned to it, and it is also reported in Table 3 under the columns titled ‘std(\( Y \))’.

From Table 3, it is clear that the proposed model forms the least number of prominent clusters compared to the other clustering methods, and still covers a fairly large (though not necessarily the largest) number of daily vectors in them. On the other hand, the average number of days per prominent cluster is highest in case of the proposed model. At the same time, this model is able to maintain homogeneity of the clusters compared to the other methods, which is indicated by the smallest standard deviation for the aggregate daily rainfall \( Y \) for days that belong to each cluster.

The clustering of the days by any method can be used to identify real-valued CRPs which are analogous to the CDPs. The CRP corresponding to each daily cluster is the mean of the DRVs of the days assigned to that cluster. The best performing value is highlighted in bold.

### Table 3. Comparison of daily cluster properties, by varying the number of clusters through \( \eta \) parameter of the proposed model

| \( \eta \) (#clusters) | #PC MRF KM SP1 SP2 | PC coverage (average) MRF KM SP1 SP2 | std(\( Y \)) 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) | 1.05 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 14 15 15         | 966 (87.8) 965 (60.3) 976 (65.1) 976 (65.1) | 1.22 2.27 1.87 |

#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.
Table 4. Comparison of daily cluster properties, by varying the number of clusters through \(\eta\) parameter of the proposed model

| \(\eta\) (\# clusters) | \(\ell_2(\phi_d)\) MRF | \(\ell_2(\phi_d)\) K-means | \(\ell_2(\phi_d)\) Spect1 | \(\text{Hamm}(\phi_d)\) MRF | \(\text{Hamm}(\phi_d)\) K-means | \(\text{Hamm}(\phi_d)\) Spect1 | \(\text{Agg}(\phi)\) MRF | \(\text{Agg}(\phi)\) K-means | \(\text{Agg}(\phi)\) Spect1 |
|------------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| 5 (146)                | 215                  | 184                  | 201                  | 31                   | 58                   | 57                   | 0.92                  | 1.05                  | 2.47                  |
| 7 (65)                 | 239                  | 211                  | 222                  | 46                   | 65                   | 65                   | 0.92                  | 1.43                  | 2.66                  |
| 8 (36)                 | 234                  | 225                  | 234                  | 53                   | 69                   | 70                   | 0.96                  | 1.47                  | 2.63                  |
| 9 (24)                 | 251                  | 232                  | 240                  | 58                   | 71                   | 72                   | 0.86                  | 1.62                  | 2.61                  |
| 10 (15)                | 255                  | 240                  | 247                  | 60                   | 73                   | 75                   | 0.86                  | 1.63                  | 2.67                  |

\(\ell_d\) is the \(\ell_2\) distance of DRVs to CRP \(\phi_d\) of corresponding cluster (equation (14)) and \(\text{Hamm}(\phi_d)\) is the mean Hamming distance of DDVs to CDP \(\phi_d\) of corresponding cluster (equation (15)). \(\text{Agg}(\phi)\) is the mean absolute error of the daily aggregate rainfall \(Y\) (equation (16)). The best performing value is highlighted in bold.

\[
\text{Agg}(\phi) = \sum_t |Y_t - \hat{\phi}_d(U(t))|.
\]  

Once again, this can be computed for CRPs identified by the proposed method, \(K\)-means and Spec1 and is shown in Table 4.

With respect to the binary patterns \(\phi^d\), Table 4 shows that those produced by the proposed method fit the DDVs better than \(K\)-means and spectral clustering (Spect2), even though the latter explicitly defines pairwise similarity measures using Hamming Distance. With respect to the \(\ell_2\) distance of DRVs from cluster centres (CRP), \(K\)-means and spectral clustering (Spect1) understandably do better than the proposed model as they explicitly aim to minimize this quantity by using \(\ell_2\) distances for objective function/pairwise similarity measure. The proposed model does not use it explicitly, but does not lag far behind. In terms of the aggregate rainfall, the clusters defined by the proposed model perform better than the \(K\)-means or Spec1.

The results in Table 4 can also be interpreted using Akaiake information criteria (AIC) for model comparison. For clustering problems, the AIC value of a model is given by \(\text{AIC} = 2K - 2 \log(P)\) where \(K\) is the number of clusters and \(P\) is the model likelihood, and a model with a smaller value of AIC is better. If we consider a Gaussian model likelihood, i.e. \(\text{X}(t) \sim N(\phi_d(U(t)), \lambda)\), then the log-likelihood is equal to the mean \(\ell_2\)-distance between each DRV \(\text{X}(t)\) and the corresponding CRP \(\phi_d(U(t))\). This is the likelihood associated with \(K\)-means, and undoubtedly it performs best. But if we consider the Hamming model likelihood, i.e. \(\prod_t \exp(\phi_d(U(t)) - \phi_d(U(t)))\) as used in the proposed model (see equation (8), then the log-likelihood is equal to the mean Hamming distance between each DDV \(Z(t)\) and the corresponding CDP \(\phi^d(U(t))\). In this case, the proposed model predictably gives the best results.

3.2. Prominent patterns

We have already introduced the term prominent cluster. The corresponding spatial patterns (CRP and CDP) will be called prominent pattern, one which appears on at least 1 day in at least 5 of the 8 years considered. Table 3 shows that the proposed method always produces about 10 such prominent patterns for a wide range of the parameter \(\eta\), and we find that these patterns are constant across different values of \(\eta\). These patterns also cover a significant number of days in this period, which is as high as 95\% for \(\eta = 9, 10\). This indicates that these 10 patterns are quite robust and frequent, and there are only a small number of days per year, which do not conform to these patterns. We also find that these days are the ones having excessive rainfall.

In the remainder of the article, we will evaluate only these prominent patterns. According to Table 3, we find that the number of prominent clusters is much higher in case of both \(K\)-means and Spec1, compared with only 10 prominent clusters in case of the proposed model. This discrepancy in the number of prominent clusters makes fair comparison of these approaches difficult. In case of both \(K\)-means and Spec1, the total number of clusters to be formed is a user-defined parameter. We therefore set this parameter so as to obtain 10 prominent clusters.

The spatial patterns we have extracted are based on only 8 years of data from 2000 to 2007. The question arises: are these patterns general enough? Can these be used to approximate DRVs from years beyond this period? Accordingly, we considered the \(\ell_2(\phi)\) and \(\text{Hamm}(\phi_d)\) measures across the period 1901–2011, i.e. the sum over time in equations (14)–(15) is taken from 1901 to 2011.
For this purpose, we run the model again using the DRVs from this period, but we use the same \( \theta \) and \( \phi \), which were estimated earlier from the period 2000 to 2007. During the inference process based on Gibbs sampling, we do not update these parameters. In effect, we estimate DDVs corresponding to the DRVs of this period and then try to approximate them with the same set of CDPs as already discovered. In case of the other approaches such as K-means and Spec1, we try to approximate these DRVs with the CRPs computed from the period 2000 to 2007. The results are shown in Table 5, and we find the prominent CRPs from the proposed model fit slightly better than those identified by the other two methods.

Finally, in Table 6, we study the spatial coherence of the CDPs and CRPs identified by different measures. This is done separately for each pattern by comparing the value at each location to those of the adjacent locations. In case of CRPs, this is measured as
\[
\sum_{k,s} \sum_{i \in \Omega(s)} \frac{|\phi_{ik} - \phi_{ik}|}{|\phi_{ik}|},
\]
while in case of CDPs, it is measured as
\[
\sum_{k,s} \sum_{i \in \Omega(s)} I(\phi_{ik} \neq \phi_{ik}).
\]
We find that the most spatially coherent CDPs are produced by the proposed method, compared with K-means, Spec2 and also EOF. Spec1 produces the most spatially coherent CRPs.

### 3.3. Graphical Representation of spatial patterns

As already discussed, each cluster produced by the proposed model and by spectral clustering with Hamming distance (Spec2) is associated with a CDP, and each cluster produced by the proposed model, by K-means, and by spectral clustering with Euclidean distance (Spec1) is associated with a CRP. In all the settings shown above, the proposed model produces about 10 prominent clusters. Corresponding to each of these, we identify 10 prominent CDPs and CRPs \( \phi_{ik} \) which are, respectively, binary and real-valued vectors of dimension \( S \) corresponding to the \( S \) locations. Thus, these vectors can be shown on a map, as done in Figure 2. Each panel corresponds to one CDP (top figure) or CRP (bottom figure), where the green locations are in state 2 (low rainfall), while blue ones are in state 1 (high rainfall).

The spatial coherence of these wet and dry zones is very notable. Clearly, some patterns are associated with pre-onset period or break spells where there is no or very little rainfall, while some others are associated with active spells with the central region (‘monsoon zone’) turning active. Some of the patterns look similar in the discrete representation, for example the seventh and ninth CDPs, but the aggregate rainfall volumes associated with them are different. Note that some locations in the north-west and south-east are not active in any of these prominent patterns. These are the regions that tend to remain dry during the monsoon. Their rare rainy days are covered by the non-prominent patterns.

To compare the CRPs obtained from our model, we also show the prominent CRPs identified by K-means and Spec1, as well as the first 10 EOF, in Figures 3–5, respectively.

We also study some properties of these clusters and the associated spatial patterns. Each of them corresponds to a distribution over daily aggregate rainfall \( Y \), and in Figure 6, we plot the mean of this distribution for each of the clusters, as \( \mu_k = \text{mean} [Y_i : U(t) = k] \). This is done for all the 24 clusters obtained by setting \( \eta = 9 \) (see Table 3), and

### Table 5. Measures of how well the spatial patterns (CRP and CDP) computed over the period 2000–07 can approximate the daily vectors (DRVs and DDVs) across the period 1901–2011

| \( \ell_2(\phi) \) | Hamm (\( \phi_d \)) | Agg (\( \phi \)) |
|----------------|-----------------|----------------|
| MRF            | K-means Spec1   | MRF            | K-means Spec2 |
| 261            | 263             | 262            | 104           |
|                 | 202             | 187            | 0.49          |
|                 |                 |                | 0.7           |
|                 |                 |                | 0.75          |

Three measures are considered: \( \ell_2(\phi) \), \( \text{Hamm}(\phi_d) \) and \( \text{AGG}(\phi) \) are defined in equations (14)–(16). The best performing value is highlighted in bold.

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

| \( spch(\phi_d) \) |
|-------------------|
| MRF               | K-means Spec2 EOF |
| 0.07              | 0.16              | 0.14             |
|                   | 0.13              |

The best performing value is highlighted in bold.
also the 10 prominent clusters among them. Also, each CDP has a fraction of the locations in ‘wet’ state $\phi_i(s) = 1$, and in Figure 6, we also plot this fraction. The plots show clearly the variation of both these quantities across the patterns, which indicates that some patterns are associated with ‘active spells’ and some with ‘break spells’.

From the assignments of $U$, we see that each CDP tends to persist for a few days, like a spell. Across the entire period, we identify such spells for each CDP and compute the mean spell length, i.e. mean number of days for which each CDP persists. The number of such spells and the mean spell length for each CDP are plotted in the rightmost panel of Figure 6. It shows us that there are more spells of dry patterns per year than spells of wet patterns, but the mean spell length is more or less uniform for all patterns.

4. Conclusion

In this article, we proposed a model that is capable of identifying the common spatial patterns of daily rainfall over India during monsoon. The model is based on advanced Bayesian nonparametric methods, which has not received
significant attention in the field of geo-sciences. The MRF model we propose allows us to identify patterns based on the data, while incorporating domain knowledge related to spatial and temporal coherence. The model creates a discrete representation of daily rainfall, which is easy to visualize and interpret.

The rainfall distribution on each day is assigned to one of these patterns, and thence the days may be grouped into clusters. We compared our method to an alternative approach used earlier: namely identification of EOF. We also performed analysis using clustering algorithms like K-means and spectral clustering and showed that our MRF-based approach produces clusters that are more homogeneous and coherent both in the data space and the geographical space. The spatial patterns we find are more representative of the daily spatial distributions of rainfall providing a better fit to data and are quite robust. The patterns we identified are interpretable to climate scientists by virtue of their spatial coherence. In addition, in our approach the data vector of each day is approximated by a single pattern unlike EOF which produces a linear combination of patterns.
We thus show that only 10 spatial patterns can represent the spatial rainfall distribution on nearly 95% days of each monsoon season with reasonable accuracy with respect to various measures (e.g. $\ell_2$-norm and Hamming similarity) compared with alternative clustering approaches, as shown in Table 5. These patterns were identified from only 8 years (2000–07), but they fit well on daily data from over a hundred years. In a companion article (Mitra et al., 2018), we will discuss the temporal characteristics of these patterns, showing that some of them to be more frequent in the early and late monsoon months (June, September), while the others are more frequent in the peak monsoon months (July, August), and also study about homogeneous zones on the landmass which are also identified by this model.

When identifying prominent patterns (i.e. those that occur in at least 5 out of 8 years, see section 3.2) using these three methods, the MRF model gives the least number of such prominent patterns which on an average are able to explain the largest number of days in the data set—the CRPs obtained by the model are much better fit to the data compared with the other methods. The number and qualitative features of the patterns are also almost constant across a range of values of the parameters that are part of the model—these CRPs are quite robust. These comparisons are discussed in detail in section 3.1. The MRF clusters are also the most coherent in real space, in the sense that adjacent positions are likely to have similar values of the discrete rainfall variable (see Table 6).

The MRF model naturally provides a clustering of the days. These clusters cover the largest fraction of the total number of days (see Table 3). The clusters we find also show the smallest mean Hamming distance of the individual cluster members to the CDPs compared with other methods, but all the methods show approximately same mean $\ell_2$
distance. Interestingly, when the CDPs identified from 2000 to 2007 data are used for clustering of the data from 1901 to 2007, the MRF clusters show the smallest mean Hamming as well as $\ell_2$ distances—the clusters we find are quite coherent in the data space, certainly when using Hamming distance which is the natural metric to use for discrete patterns.

**Declaration**

Funding: none.
Ethical approval: none.
Conflict of interest: none.

**Acknowledgements**

A.M. was with ICTS-TIFR, Bangalore, India, when most of this work was done. A.M., A.A. and S.V. acknowledge the support of the Airbus Group Corporate Foundation Chair in Mathematics of Complex Systems established in ICTS-TIFR and TIFR-CAM. A.M. and A.A. thank the Statistical and Applied Mathematical Sciences Institute (SAMSI), Durham, NC, where a part of the work was completed.

**References**

Fu Q, Banerjee A, Liess S, Snyder PK. Drought detection of the last century: an MRF-based approach. In: Proceedings of the 2012 SIAM International Conference on Data Mining. SIAM, 2012, pp. 24–34. [https://experts.umn.edu/en/publications/drought-detection-of-the-last-century-an-mrf-based-approach](https://experts.umn.edu/en/publications/drought-detection-of-the-last-century-an-mrf-based-approach)

Gadgil S. The Indian monsoon and its variability. *Ann Rev Earth Planet Sci* 2003; 31: 429–67.

Gadgil S, Gadgil S. The Indian monsoon, GDP and agriculture. *Econ Pol Week* 2006; 41: 4887–95.

Gadgil S, Joseph PV. On breaks of the Indian monsoon. *J Earth Syst Sci* 2003; 112: 529–58.

Gadgil S, Narayana Iyengar R. Cluster analysis of rainfall stations of the Indian peninsula. *Q J R Meteorol Soc* 1980; 106: 873–86.

Gadgil S, Rajeevan M, Nanjundiah R. Monsoon prediction – why yet another failure. *Curr Sci* 2005; 88: 1389–400.

Ghosh S, Mallick BK. A hierarchical Bayesian spatio-temporal model for extreme precipitation events. *Environmetrics* 2011; 22: 192–204.

Goswami BN, Ajaya Mohan RS. Intraseasonal oscillations and interannual variability of the Indian summer monsoon. *J Clim* 2001; 14: 1180–98.

Greene AM, Robertson AW, Kirshner S. Analysis of Indian monsoon daily rainfall on subseasonal to multidecadal time-scales using a hidden Markov model. *Q J R Meteorol Soc* 2008; 134: 875–87.

Greene AM, Robertson AW, Smyth P, Trigilia S. Downscaling projections of Indian monsoon rainfall using a non-homogeneous hidden Markov model. *Q J R Meteorol Soc* 2011; 137: 347–59.

Hartigan JA, Wong MA. Algorithm as 136: a k-means clustering algorithm. *J R Stat Soc C (Appl Stat)* 1979; 28: 100–8.

Haylock M, McBride J. Spatial coherence and predictability of Indonesian wet season rainfall. *J Clim* 2001; 14: 3882–7.

Holtsclaw T, Greene AM, Robertson AW, Smyth P. Bayesian nonhomogeneous Markov models via pòlya-gamma data augmentation with applications to rainfall modeling. *Ann Appl Stat* 2017; 11: 393–426.

Kindermann R, Snell L. *Markov Random Fields and Their Applications*. Providence, RI: American Mathematical Society, 1980.

Krishnamurthy V, Shukla J. Intraseasonal and seasonally persisting patterns of Indian monsoon rainfall. *J Clim* 2007; 20: 3–20.

Krishnan R, Zhang C, Sugi M. Dynamics of breaks in the Indian summer monsoon. *J Atmos Sci* 2000; 57: 1354–72.

Kulkarni S, Deo MC, Ghosh S. Impact of active and break wind spells on the demand–supply balance in wind energy in India. *Meteorol Atmos Phys* 2018; 130: 1–17.

Lavine M, Lozier S. A Markov random field spatio-temporal analysis of ocean temperature. *Environ Ecol Stat* 1999; 6: 249–73.

Mitra A, Apte A, Givindarajan R, Vasan V, Vadlamani S. *Spatio-temporal Patterns of Indian Monsoon Rainfall*. 2018. Accepted for publication in Dynamics and Statistics of the Climate System.

Moron V, Robertson AW, Ghil M. Impact of the modulated annual cycle and intraseasonal oscillation on daily-to-interannual rainfall variability across monsoonal India. *Clim Dyn* 2012; 38: 2409–35.

Moron V, Robertson AW, Pai DS. On the spatial coherence of sub-seasonal to seasonal Indian rainfall anomalies. *Clim Dyn* 2017; 49: 1–21.

Moron V, Robertson AW, Qian J-H, Ghil M. Weather types across the maritime continent: from the diurnal cycle to interannual variations. *Front Environ Sci* 2015; 2: 65.

Moron V, Robertson AW, Ward MN, Camberlin P. Spatial coherence of tropical rainfall at the regional scale. *J Clim* 2007; 20: 5244–63.
Neal RM. *Probabilistic Inference Using Markov Chain Monte Carlo Methods*. Technical report, Department of Computer Science, University of Toronto, 1993.

Ng AY, Jordan MI, Weiss Y. On spectral clustering: analysis and an algorithm. In: *Advances in Neural Information Processing Systems*. Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Cambridge, MA: MIT Press, 2002, pp. 849–56. [http://dl.acm.org/citation.cfm?id=2980539.2980649](http://dl.acm.org/citation.cfm?id=2980539.2980649)

Pai DS, Rajeevan M. *Indian Summer Monsoon Onset: Variability and Prediction*. National Climate Centre, Indian Meteorological Department, 2007.

Pitman J. Exchangeable and partially exchangeable random partitions. *Probab Theory Rel Fields* 1995; 102: 145–58.

Rajeevan M, Bhave J, Kale JD, Lal B. High resolution daily gridded rainfall data for the Indian region: analysis of break and active. *Curr Sci* 2006; 91: 296–306.

Rajeevan M, Gadgil S, Bhave J. Active and break spells of the Indian summer monsoon. *J Earth Syst Sci* 2010; 119: 229–47.

Ramamurthy K. Monsoon of India: some aspects of the ‘break’ in the Indian southwest monsoon during July and August. *Forecast Man* 1969; 1: 1–57.

Saha M, Mitra P, Nanjundiah RS. Autoencoder-based identification of predictors of Indian monsoon. *Meteorol Atmos Phys* 2016; 128: 613–28.

Sang H, Gelfand AE. Hierarchical modeling for extreme values observed over space and time. *Environ Ecol Stat* 2009; 16: 407–26.

Sikka DR, Gadgil S. On the maximum cloud zone and the ITCZ over Indian, longitudes during the southwest monsoon. *Mon Weather Rev* 1980; 108: 1840–53.

Singh N, Ranade A. The wet and dry spells across India during 1951–2007. *J Hydrometeorol* 2010; 11: 26–45.

Suhas E, Neena JM, Goswami BN. An Indian monsoon intraseasonal oscillations (MISO) index for real time monitoring and forecast verification. *Clim Dyn* 2013; 40: 2605–16.