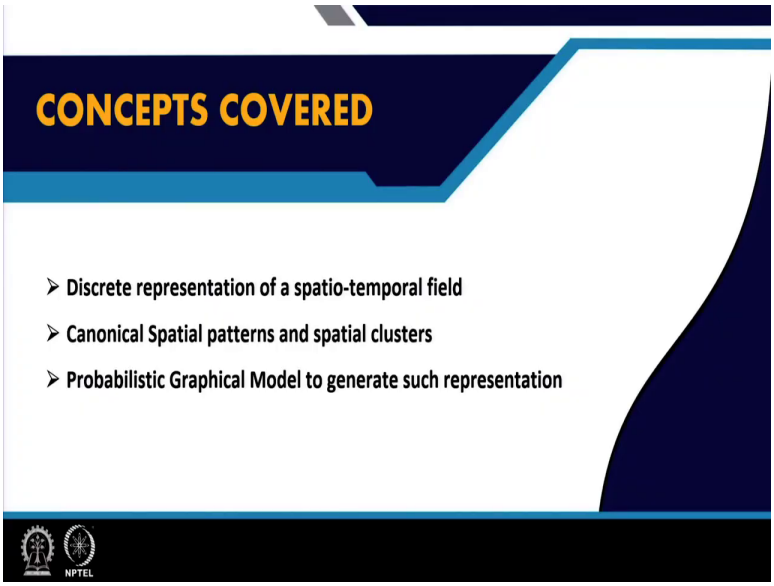
Machine Learning for Earth System Sciences
Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 03
Machine Learning for Discovering New Insights
Lecture - 25
Discovering Clustered Weather Patterns

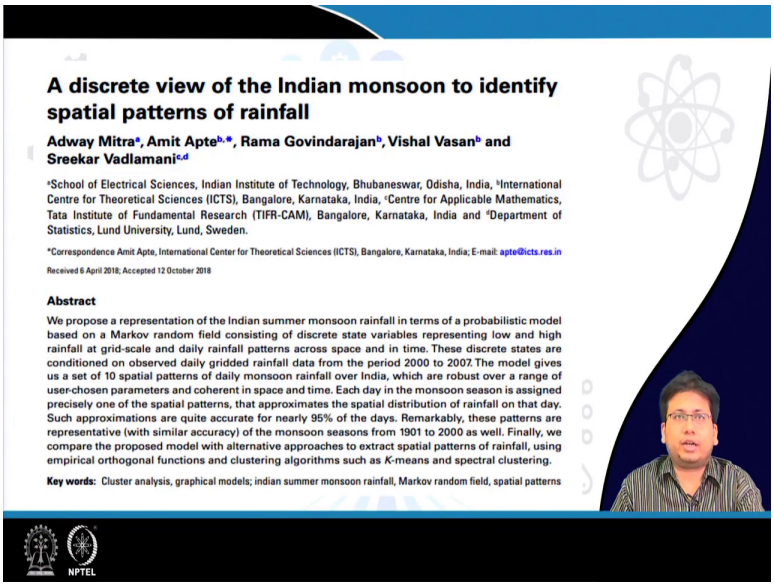
Hello, welcome to lecture 25 of this course on Machine Learning for Earth System Science. This lecture is about Discovering Clustered Weather Patterns using graphical models. So, we are still in module 3, where we are using Machine Learning for discovering various insights related to the earth system sciences.

(Refer Slide Time: 00:45)



So, the concepts we are going to cover in this lecture are – firstly, how to create a discrete representation of a spatio-temporal field of any geophysical variable; secondly, how to identify canonical spatial and temporal patterns and spatial clusters out of such a field and thirdly, how probabilistic graphical models can be used to generate like such a representation or such patterns and so on.

(Refer Slide Time: 01:11)



So, the we will start discussing this paper a discrete view of Indian monsoon to identify spatial patterns of rainfall. So, like Indian monsoon is something we have discovered earlier also. The broad problem which we are interested in today is like we what we have discussed earlier is Indian monsoon season has a lot of spatio temporal variability; that is, there is intra seasonal variability and inter seasonal variability.

When we are considering one monsoon season then even within that season the rainfall is not spatially this uniform all through like for all the 122 days from 1st June to 30th September. Rather like on there are certain periods where the rainfall is mostly concentrated in one region, in another period the rainfall is concentrated in another region, some days are mostly dry and so on.

So, our aim here is to like identify certain broad spatial patterns along which the rainfall may be distributed on any given day of our typical monsoon season. So, like coming to this paper, we propose a representation of the Indian summer monsoon rainfall in terms of a probabilistic model based on Markov random field consisting of discrete state variables representing low and high rainfall at grid scale and daily rainfall patterns across space and in time.

These discrete states are conditioned on observed daily graded rainfall data from the period 2000 to 2007. The model gives us a set of 10 spatial patterns of daily monsoon rainfall over India,

which are robust over a range of user chosen parameters and coherent in space and time. What I mean by coherent in space and time? So, coherent means like the patterns which we are talking about, we will see those patterns shortly like they are like spatially contiguous that is you would not see that one region has rainfall, another region is like one another adjacent region is dry and so on.

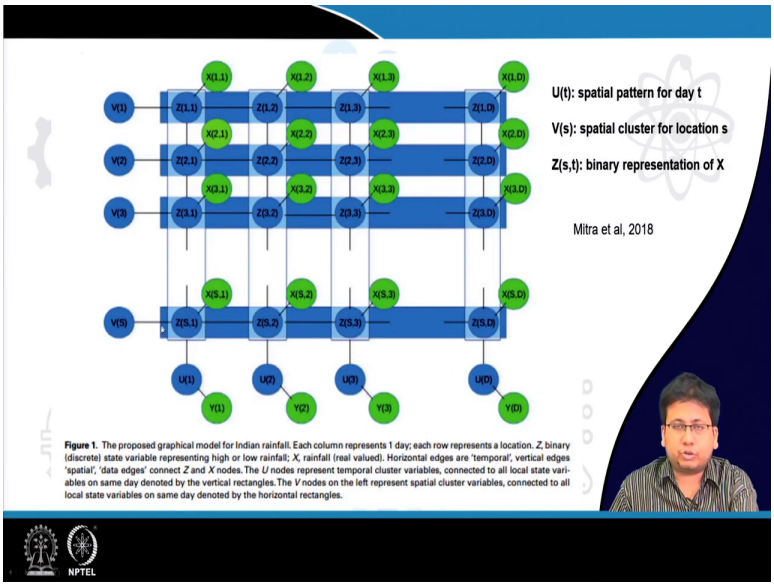
You want it there is to say it only patchy, but like we will be able to identify clear blocks where the rainfall activities are concentrated and when I say coherent in time it means that one of these patterns it perceives for a few days. It is not like that it is one day it is like this, the next day it will be something else. It is not like that. There is some kind of temporal stability or continuity in the process.

So, the like these are; like these are like in sync with the various understandings or the various conventional wisdom or related of climate scientists related to Indian monsoon where like. So, there like there are concepts like MISO – Monsoon Intra Seasonal Oscillations or active and break spells like these are all concepts that have been studied by climate scientists and Indian meteorologists related to the Indian monsoon.

This paper or the or the papers we are going to discuss today try to bring out those same concepts in a more like comprehensive way or easy to like in ways that are easier to visualize. Each day in a monsoon season is assigned to precisely one of the spatial patterns that approximates the spatial distribution of rainfall on that day. Such approximations are quite accurate for nearly 95 percent of the days.

Remarkably, these patterns are representative of the monsoon seasons from 1901 to 2000 as well. Finally, we compare the proposed model with alternative approaches to extract spatial patterns of rainfall, using empirical orthogonal functions and clustering algorithms such as K-means and spectral clustering.

(Refer Slide Time: 05:17)



So, this is the probabilistic graphical model which we are talking about. You may remember that in an earlier lecture about the anomaly detection we had all this discussed a similar graphical model like this. So, these are the observations the like we may call this as X. So, in this case it is the rainfall. So, along each row we have the different the days and the along each column we have the different locations.

So, two locations are like are connected by an edge if they are adjacent to each other. I mean like the way it is drawn here it might seem that each location has only two neighboring locations, but that is actually not the case. So, like when we actually plot all these locations on a like on the map. So, each of these locations represents a grid point. So, each grid will have maybe four neighbors on all on the four sides. So, or it can have more neighbors also if you consider the diagonals.

So, these there can be more edges or more neighboring nodes also which we have not shown in this model, but these vertical edges basically indicate the spatial connections. And, the horizontal edges these indicate temporal connections in the sense that these are edges a successive days. This is day 1, this is day 2 etcetera and like this kind of a structure like it can like continue for like at the entire season or something like that.

So, for a so, these green nodes these are the locations that is $X(s,t)$ in general will mean the rainfall at location s on day t and corresponding to each of these observations we have the discrete or binary representation which is indicated by Z. So, this binary variable it will have two values; like one – meaning that there is significant rainfall, two – meaning that there is less rainfall or no rainfall.

Now, how much is significant etcetera these are like there is no hard threshold on this instead it is like left as like there you can say that it is something like a mixture model there is a distribution for probability distribution for dry days and there are different probability distribution for wet days. So, on any given days rainfall like we will have to see which distribution it fits better.

And, that and, in doing so, we also have to take into account the neighboring Z values that is both the spatial neighbors as well as the temporal neighbors. So, if there is like an isolated rainfall event which is like very high in that case Z may turn out to be 1, but if there if there is a like a mild rainfall event or maybe like 2 millimeters of rainfall happens in a particular place.

Even though its surrounding regions receive no such rainfall and or that that same place in the previous and next days also does not receive any next rainfall, then that 2 millimeter of rainfall may just be ignored and like it on that day that location may be considered to be in a dry state. Now, we additionally we have these U variables and the V variables. The U variables indicate the spatial pattern of day t.

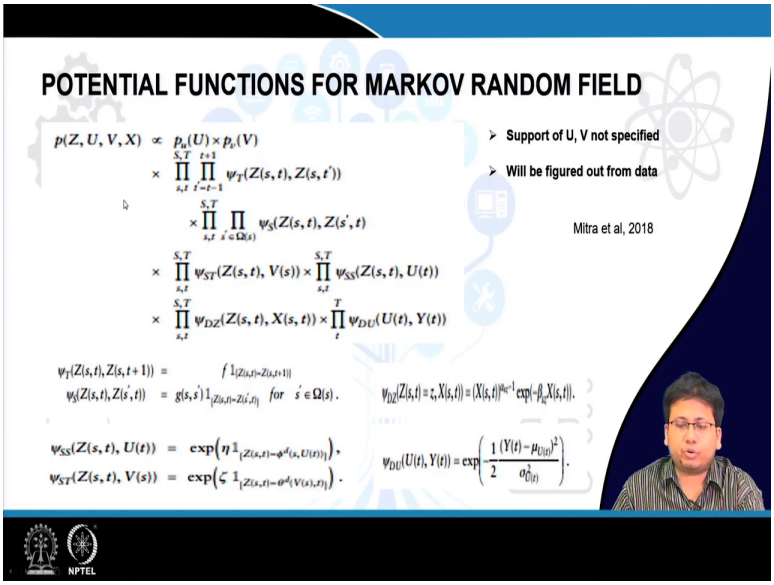
Now, as I said that there are us like we want to identify a few spatial patterns such that the rainfall distribution on any given day as obtained from the these X and Z values they can be approximated by any one of those patterns. So, for this particular day there will be one pattern, for this day there will be another pattern and so on. So, which day belongs to which pattern that is encoded by this variable U.

Now, similarly the different locations there are let us say there are S number of locations. So, they are like these locations again can be grouped into some we can say homogeneous regions which have very similar temporal patterns of rainfall that is whenever the rainfall is high in one

location, in the other locations also in the same region it will be high. If rainfall is low in on a any given day rainfall is low in a location in the other locations also it will be low and so on.

So, that way like all the locations basically can be divided into clusters and that cluster membership of each location is indicated by these variables.

(Refer Slide Time: 10:07)



So, an important thing to note here is that the support of U and V is it is not specified. I mean, obviously, U cannot be more than D that is the total number of days. Worst case every single day we will have it like will be an unitary cluster having its own pattern and similarly, worst case all the locations will be 1, 1 cluster each. So, maximum value of V is of course, S and for U is D, but we will not consider like of course, that is not a desirable scenario.

But, exactly what values U and V should take that we do not specify unlike say something like k-means clustering we do like the number of clusters is not specified rather it needs to be found out from the data itself that is up to the model to do so. That is why this class of models are known as the non-parametric models.

So, in this like obviously, like as we do in such situations we imagine these observations X to be realizations of a random variables and these others Z, U, V these are also random variables which are unknown or which we do not observe. So, we define the joint distribution of all these

variables in like in the following way. So, first of all we have the distributions on the U variables, then on the V variables and then like we have all the edge potentials as happens in a Markov random field.

So, like the. So, these are like the temporal or the edges or the I mean the horizontal edges, so, they have some potential functions. Similarly, there are these vertical edges between two adjacent locations they have their potential functions. Then like then there are all these look at these edges which are connecting the Z variables to the corresponding spatial cluster variables, then there are the edges like connecting the each Z value to our Z each Z variable to the corresponding like pattern variable U.

So, here like all those edges have not been individually shown, but you can the way it has like. So, these horizontal edges as you can see they are connected to this whole thing; that means, that this $V(1)$ this is actually connected to all the Z variables on this day. Similarly, this $V(s)$ this is all sorry, at this location. Similarly, this $V(s)$ this node is connected to all the variables in this location and so on.

Similarly, this $U(2)$ this is connected to all the locations on this day and so on and so forth. So, we have like. So, these all these edges are there so, we have these edge potentials. Now, these edge potential functions are defined in such a way as to like as to promote spatiotemporal coherence.

So, if you remember in the earlier paper about anomaly detection, we have we had a talked about defining the edge potentials in such a way that it will the potential function will take a high value if the like if the nodes if the end to end nodes they have equal values otherwise it will take a low value.

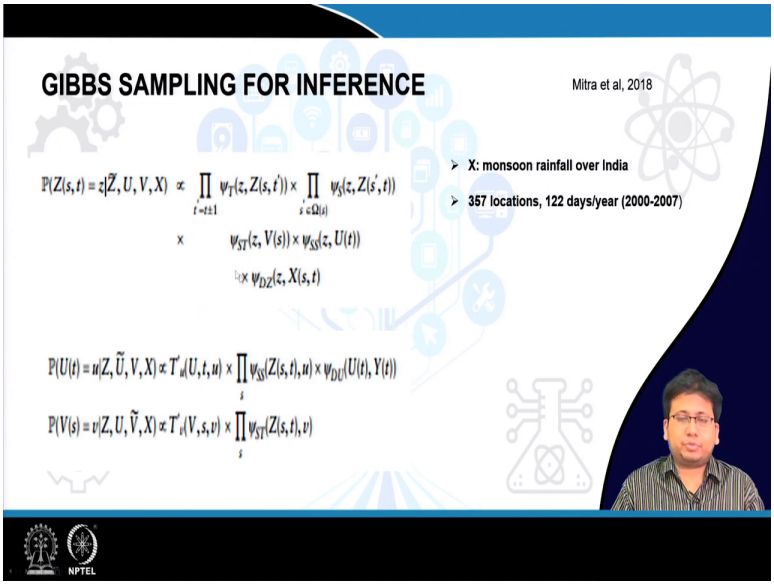
So, that like the there is an incentive for the two like end variables that took very the nodes that are being connected to have if there is some incentive for them to have the like equal values there and that is what like promotes the property of spatial and temporal coherence. Similarly, on the edge potential functions which connect the hidden nodes to the observation nodes that is Z to X variables they also have like we like associate these we represent each of these X variables as some kind of like mixture of gamma distributions.

So, for as I said for at each location we assume that Z can take two very well two values, 1 and 2; 1 means like wet day and 2 means a dry day. So, for 1 that is wet days there is 1 probability distribution on X and for dry days also there is a different probability distribution on the variable X. Both of those will be gamma distributions, but with different parameters.

So, apart from that, similarly the Y is like the total amount of rainfall on any given day so, that is associated with the spatial patterns. So, like we also assume some kind of like we assume a Gaussian distribution on Y, the total amount of rainfall over the individual like over the entire landmass. So, if you consider a particular location the rainfall is non-Gaussian is highly non-Gaussian and it is best represented by something like gamma.

But, if you consider the total aggregate rainfall over the entire region then the Gaussian is a good distribution to represent it. That is why from the U to Y edge potential we like use the Gaussian pdf as the potential function.

(Refer Slide Time: 15:32)



So, once we have this like once we have established all these potential functions, the next task is to estimate the values of the variables that we do not know that namely Z, U and V and as is the norm in this kind of situations we in like we estimate then by the process of sampling that is to say we like have some initial values of Z, U and V.

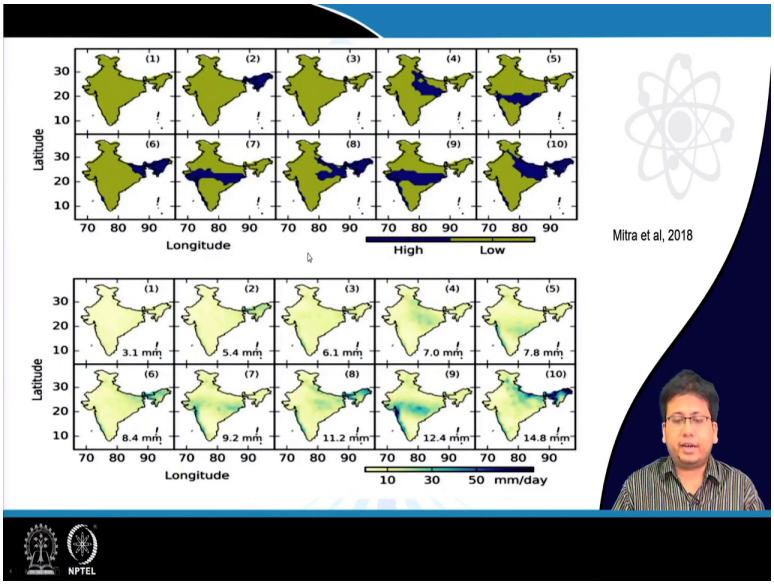
Then, we repeatedly sample new values of each variable keeping the other all the other variables as constant. So, for that we need conditional distributions for every single variable every single latent variable and these are the those distributions. So, like this, is of course like a binary variable. So, this will be something like a Bernoulli distribution, but what is the parameter of that? So, for that we need to like calculate all these things.

Similarly, this U, V like both are of these are discrete variables. So, they will have some discrete distributions. So, what, but what will be the values of each probability of So, like that can be calculated using the like all these potential functions. So, like basically these conditional distributions come from the joint distribution that has been calculated here.

So, now will I so, as is the usual process of Gibbs sampling we keep draw keep on drawing samples of all these and keep storing them and finally, like when the thing has converged we like check the samples of every variable and like we can take the mode value of those samples. So, that will be something like the map estimate of the all these latent variables. So, this thing is this data is now applied to the Indian region.

So, we consider 357 locations and there are one like 122 days in every year. So, we consider 8 years in like successively 2000 to 2007.

(Refer Slide Time: 17:36)



So, now the so, the claim is the of this work is that like if you consider the rainfall map over India on any given day on any given day it can be expressed like it is like roughly similar to one of these ten patterns. That is to say, like these ten patterns are enough to represent almost all days of the Indian monsoon; of course, not completely, but approximately.

So, like these 10 clusters which we have found these are what we call as prominent clusters or prominent pattern that is such a pattern which is present on several days of every given year. That is that is these patterns we have found are not from one single year, but across the period of like 8 years that we have considered 2000 to 2007. Not only that also although we have extracted them from this 8 years.

But, in fact, if you consider years beyond this range also I mean after 2007 or before 2000, then also any the same claim holds that is any days rainfall map is approximately one of these 10. And so, the question we may be interested to ask is how many days are actually covered by these prominent patterns. And, the answer turns out to be above like nearly 93.8 percent, it can be so high.

Now, the number of these patterns which we are getting how are we getting that that is how did we get exactly 10 patterns? So, you might remember that we had not specified U and V. So, U means what? U with the support of U is basically the number of patterns that we are getting.

So, although we had not specified beforehand to the model that we want 10 patterns only or something like that these are the 10 like we in fact, we as I mentioned earlier this is left very as a variable which can be as high as like D itself but, we are in, but we are getting only about 10 or 11 patterns. In fact, we are that is not completely true.

We get more patterns than this, but many of them are like a rare patterns which appear rarely maybe once in a year or something like that, but these are the prominent patterns which appear as in every year and for a minimum number of days. And, so, like we see that we have from the like when we use this the model that we just discussed the graphical model that we just discussed, we get like 10 to 11 patterns like there is one para well there is one parameter in the model which we may vary to increase or decrease the number of clusters we get, but like that, but we always get this number of prominent patterns.

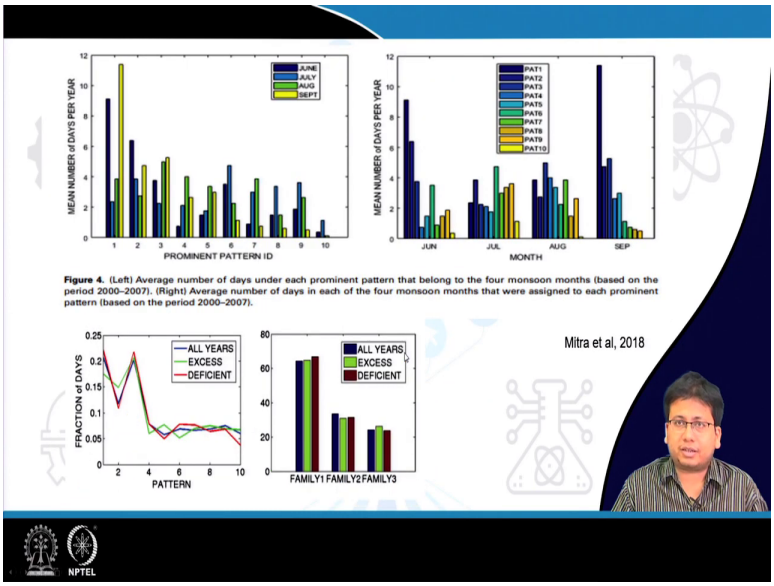
Now, if we use some other clustering techniques suppose instead of going for the model that we considered we just discussed instead of that suppose if we go for k-means clustering or spectral clustering or something like that, then the number like we do not get so clear and prominent patterns like this. In fact, like we get much smaller and less coherent clusters.

So, then we can study once we have all these clusters we or these patterns whatever you can call them, we can always study various properties of it such as like as you can see that these different patterns they have different they are associated with different volumes of rainfall. Some patterns are drier, some are more wet. And, then also so, like we can like study the these properties of the different patterns.

Also, we can study that is what is the mean aggregate rainfall in each of the patterns the spells of each patterns per year. So, like as so, these patterns actually show as I was telling earlier some amount of temporal coherence, that is, if day t belongs to one of these patterns most likely day $t + 1$ will also belong to the same pattern. It would not change abruptly from day to day.

So, there what will happen is there will be spells each like once one pattern is reached it will continue for maybe 3 or 4 days and so on. So, what we have studied also these papers do is they study the mean spell lengths of these patterns. So, like 2 days, 3 days whatever.

(Refer Slide Time: 24:22)



And, we are also the each of these patterns repeat several times a year and it also happens that in some months some patterns are more common and in some other months they are less common. So, for example, the these dry these patterns 1 and 3 which are mostly dry these happen mostly in the either in June or in September; that is in June before the onset of monsoon and in September after the withdrawal of monsoon, but in July and August also they can happen. So, those are times are known as the break spells.

So, Indian meteorologist define something like so, break spells in Indian monsoon which are short phases of a few days where the rainfall activity stops and almost all parts of the country do not receive rainfall. Similarly, there are active spells also. So, like we see that some of these patterns like 1 and 3 these are actually associated with the those break spells and some of these like patterns like a 9 or 10 which where the aggregate amount of rainfall is very high.

So, these are typically associated with the active spells and the this is like this fact this fact is actually demonstrated by the like if we see the month wise distribution of the different patterns ah. So, like so, we for example, we know that June and September they often have less rainfall because in June there is the rainfalls do not start over the whole country until the monsoon onset happens and similarly, in September the rainfall stops after the monsoon withdrawals from different parts of the country.

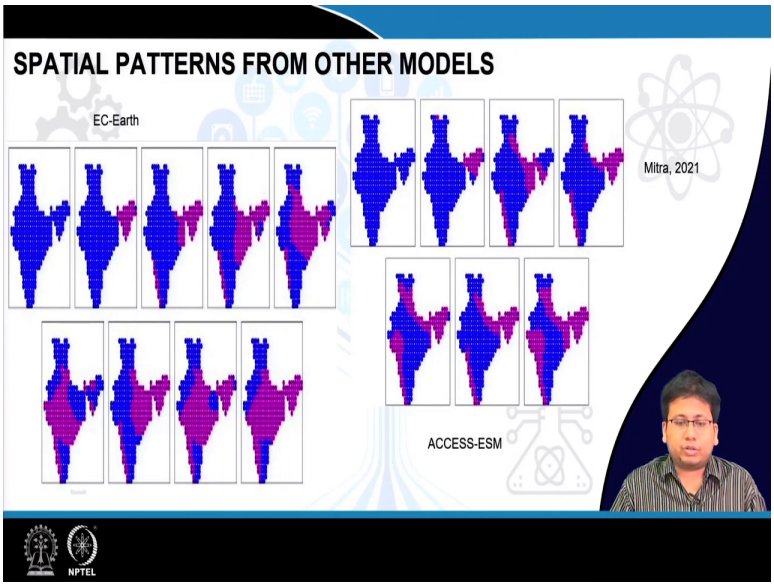
Now, there is this these concepts of these patterns these are also related to the concept of inter like inter annual or inter seasonal variability of Indian monsoon. So, it is known that there are some years of excess rainfall, some years of deficient rainfall etcetera. So, in years of excess rainfall it is found that some patterns especially like the more these patterns they are more frequent while in the deficient years of years of deficient rainfall, some other patterns might are found to be more prominent.

So, in fact, these patterns can be divided into different families. So, as you can see like the one there is one family where the rainfall is mostly in is like either absent or restricted to the coastal regions and the Northeast. Then there are some other there is another family like this for patterns 4, 6, 8 and 10 where the rainfall is mostly in the Indo-Gangetic plain and in the Northeast and, then there is a family 3 where the rainfall is mostly in the Central India.

So, we see that there is a like a difference in the distribution of these families in the like in the excess years and the deficient years and so, like in like especially in the excess years one family like might dominate while in other years the that the that family may be less dominant.

Or so, like especially the Central India pattern this is found to the Central India family these are found to be more frequent in the years of excess rainfall, while in the years of deficient rainfall this family is found to be more dominant or more frequent.

(Refer Slide Time: 28:03)



Now, these the patterns which we just discussed these are the patterns which we have got obtained from the observational data from India from the India Meteorological Department, but we also have all the so many process models, the global climate models and so on which aim to simulate climate all over the world including the monsoon of India.

So, like one question of which India which climate scientists have been asking over the years is how accurately do these models simulate the various phenomena all over the world like including say or especially Indian monsoon. And, like till recently the common refrain among scientists was that Indian monsoon is not represented very well by these models.

So, what we try to do is we actually looked at the Indian monsoon simulations by these models and try to bring out the patterns and we try to see this particular aspect the spatio-temporal

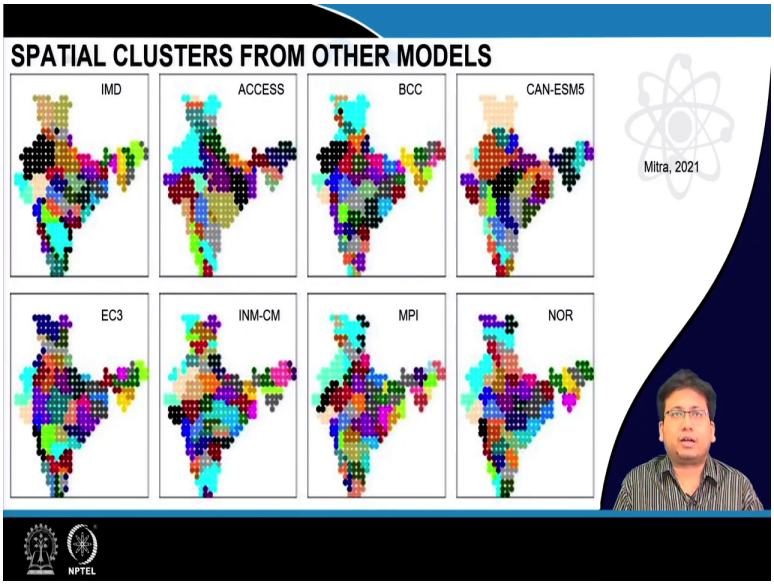
distribution of rainfall over India how like how well the Indian how well all these climate models are able to represent them and this is what we see.

So, we so, these are the like. So, EC-earth this is one GCM this access the access earth system model, this is another one, this is designed by some agency in Australia, this EC this is by the European organizations and so on. So, these are like. So, these are the spatial patterns which we get from the observational data.

But, the if we look at the simulation data then these are the spatial patterns we get which seem to have some relations, but are largely different. Like we see that like in this case we see that there are only seven patterns and most of these show rainfall like I either along the these Northeastern and coastal regions or like we see some dry entirely dry days, but the Central Indian patterns they are absent.

On the other hand, if you see these we see like here also we see that these patterns look rather different from the patterns we obtained there. So, that suggests that these model most of these models even if they can represent the aggregate amount of rainfall in India they are you are usually unable to represent or simulate the spatial distribution of rainfall over the landmass on any given day.

(Refer Slide Time: 30:50)

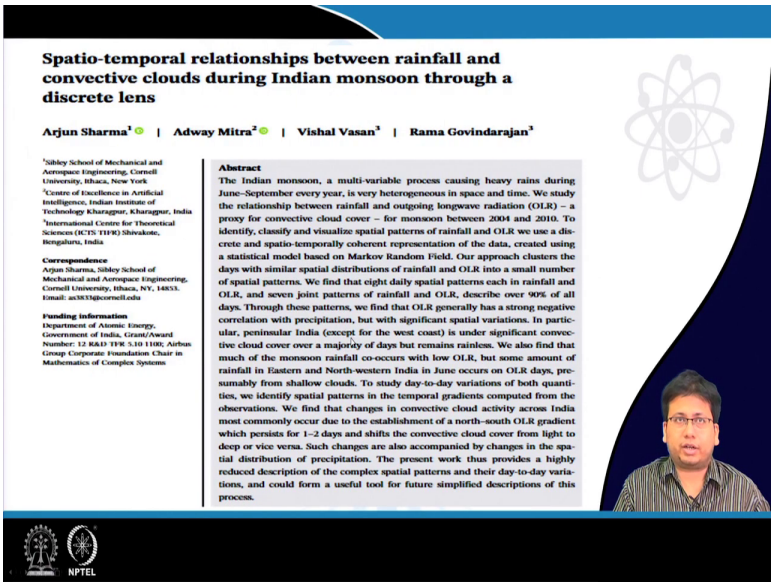


And, like you remember that along with the U variables we had also mentioned the V variables which denote the spatial clusters. So, basically they divide the region the whole landmass into small homogeneous regions which have almost similar rainfall behavior. So, so, what are those regions? So, which are those regions?

So, here the these colors indicate the various regions that like that we have that have been found on the basis of the V variables. So, we can say that each of these regions like they like they might be some local regions which have similar topography characteristics and so on. So, these regions they have like if you consider any two locations within the each of these regions or clusters you will see like the rainfall time series is like more or less same.

So, then the same question we can ask that like from each of the different models which simulate Indian monsoon like do we get the same set of regions or I mean homogeneous regions or not. So, we can like on that we can find out on the basis of the V variable.

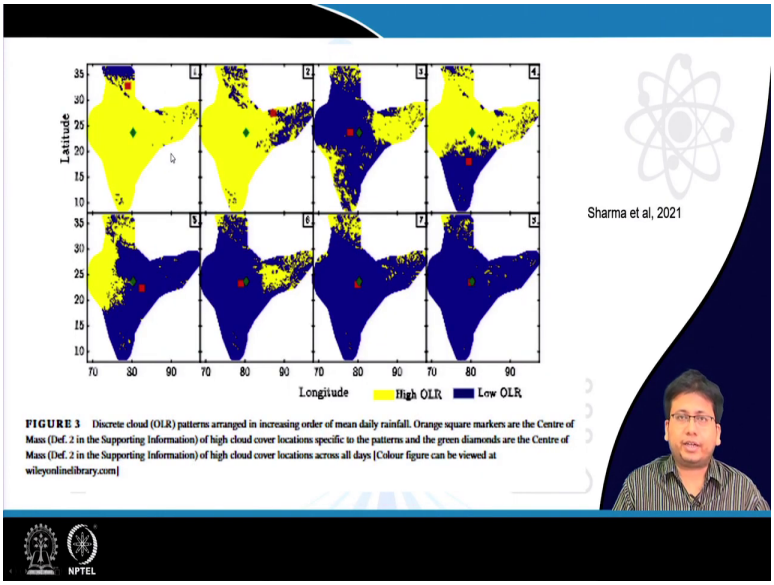
(Refer Slide Time: 32:14)



And, we find that roughly there are some broad similarities, but overall it is not like very great and now, this kind of analysis which we have done of creating a discrete representation, this did not be done for one variable but it can be done for multiple variables at the same time. So, for

example, in this paper like two variables are considered the rainfall as well as the convective cloud cover.

(Refer Slide Time: 32:35)



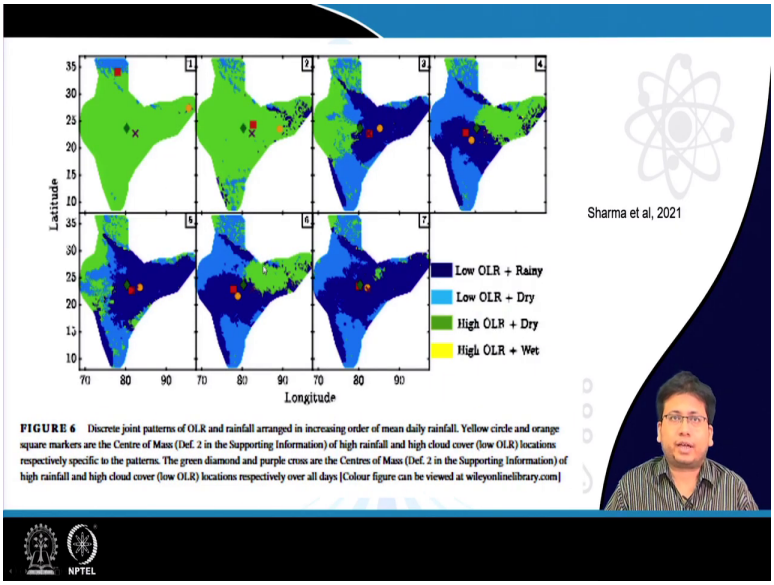
So, cloud convective cloud cover is cannot be of course, measured directly, but there is this outgoing long wave radiation which is obtained like from or which is measured by satellites. So, the general idea is that the when the earth when the some region of the earth is covered by cloud it is unable to radiate much. So, if the OLR is low that indicates that there is a; there is a there is probably a convective cloud cover over that region.

On the other, hand if the OLR is high that is the satellite is able to receive a lot of radiation from the earth; that means, the skies are clear that is there are no clouds. So, this OLR this is considered as a plot proxy for convective cloud. Now, considering this as a variable again we can construct the similar kinds of spatial patterns as we discussed earlier and these are the some of the spatial the eight spatial patterns of OLR distribution over India over the Indian landmass during the rainy season.

So, we find that there are some days where most of the country has like high OLR that is very low cloud cover and there are also days where most of the region is entirely having low OLR that

is there is there is widespread cloud cover. While there are some days where the South has low OLR, but the North does not have and like vice versa as well.

(Refer Slide Time: 34:05)

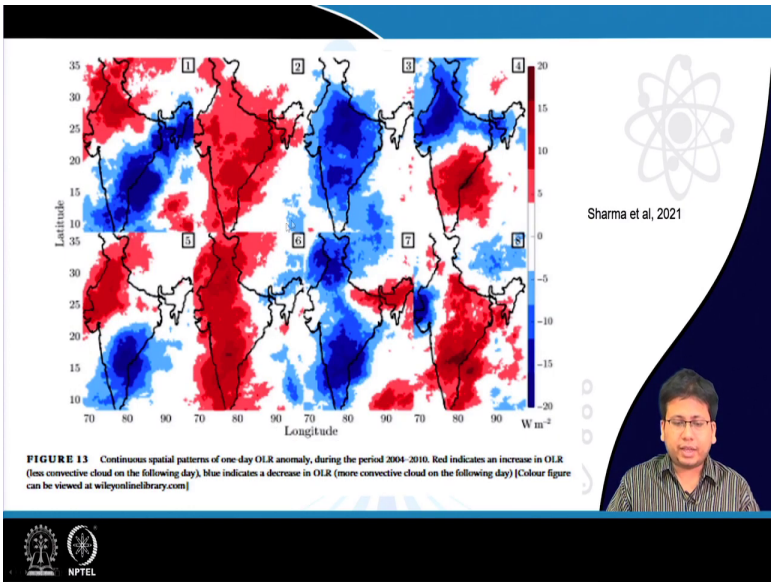


And, and now these patterns can be like be jointly in terms of the OLR and precipitation and so, we can find like basically instead of having two states like this. So, the Z variables which we have talked earlier they are by they have been binary so far that is wet or dry low OLR high OLR and so on, but like we can make them like four values.

So, like as you can see in these cases like so, dark blue means low OLR plus rainy that is there is strong convective cloud plus there is rainfall. Then, there is the situation where there is a low OLR that is convective clouds are there, but there is not rainfall. So, as you can see the southeastern part of India usually does not receive so much of rainfall during the monsoon season, but it can have like heavy cloud cover.

This happens mostly due to the orography of the region with the presence of mountain and so on. Similarly, there are this is the situation where there is a high OLR that is no clouds and there is obviously, no and no rainfall also.

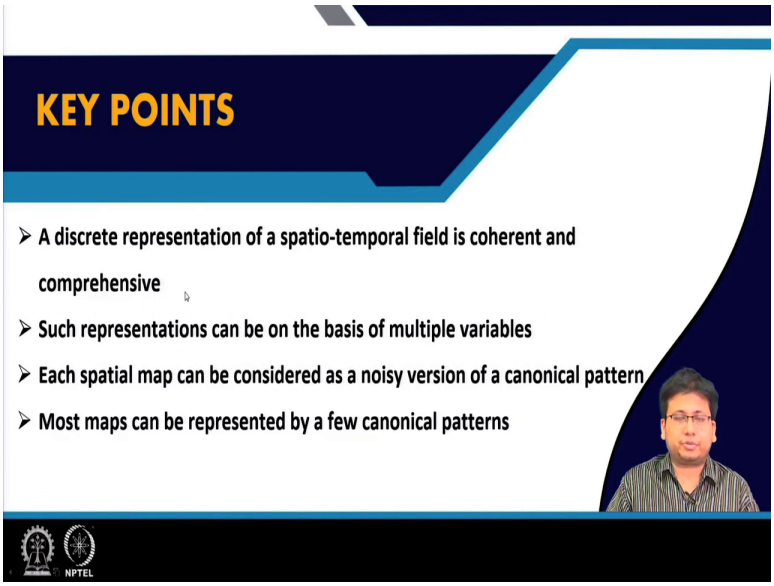
(Refer Slide Time: 35:14)



And, similarly the movements of the cloud bands from one region to another can also be studied using this kind of representation.

(Refer Slide Time: 35:21)

(Refer Slide Time: 35:25)



So, these are the references of the different papers which we discussed today. So, the key points to take away from this lecture is that firstly – a discrete representation of a spatio-temporal field is coherent and comprehensive that is by looking at the maps that are created we can understand, we can make some sense of where the rainfall is located and so on.

Such representations can be the basis of multiple variables and each as I said like for example, it both on precipitation on OLR, apart from that we can add other variables also maybe wind directions and various other things. Now, each spatial map can be considered as a noisy version of each canonical pattern. If you consider the spatial map of any variable on any given day it is going to roughly represent or roughly resemble any one of the pattern that have been identified in the way. And, finally, most of the maps can be represented in this way.

So, that brings us to the end of this lecture. So, we will continue in the next lecture again.

Thank you.