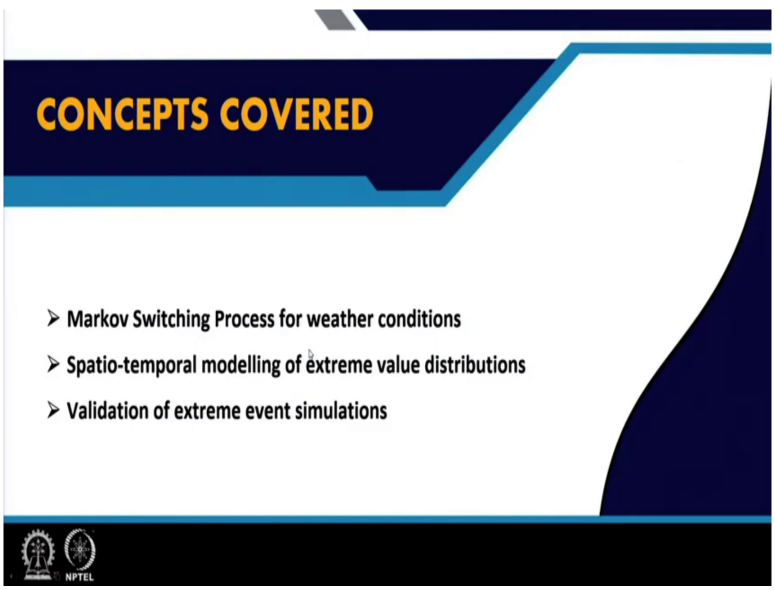
Machine Learning for Earth System Sciences
Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 03
Machine Learning for Discovering New Insights
Lecture - 21
Spatio-Temporal Modelling of Extremes

Hello everyone, welcome to lecture 21 of this course on Machine Learning for Earth System Science. So, today we will be talking about Spatio Temporal Modelling of Extremes. So, we are still in module 3 where we are like exploring different applications of Machine Learning for Discovering New Insights in earth systems.

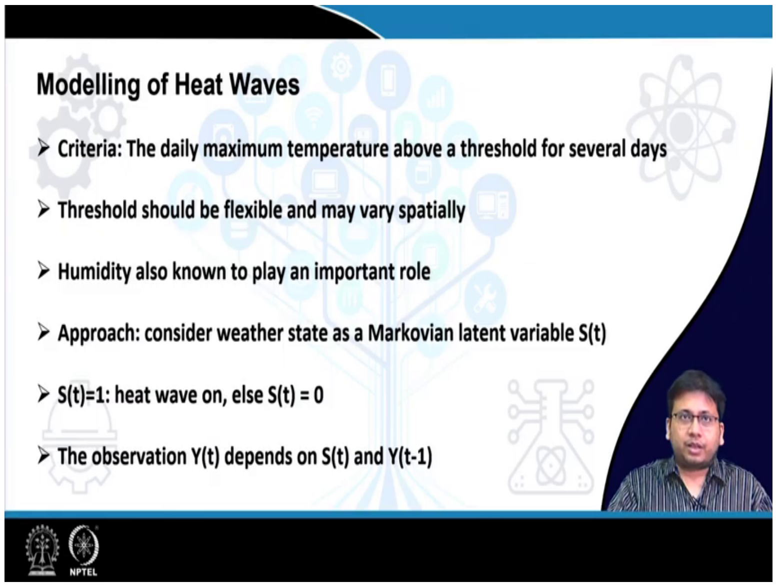
(Refer Slide Time: 00:46)



So, the concepts which we are going to cover to in this lecture are Markov Switching Process for weather conditions, Spatio temporal modelling for extreme value distributions and validation of extreme event simulations.

So, like you like. So, these are the most of these concepts we have studied in isolation in module 1 to in this lecture we will see some concrete uses of this or some we like we will come across some research papers where these concepts have been used for like for various applications in the in earth sciences.

(Refer Slide Time: 01:19)



So, the first topic which we will talk about today is the modelling of heat waves. So, we know a heat wave is like a is a phenomena where the temperature is high for a in a region for several days at a stretch and by when I say it remains high it means, above some particular threshold.

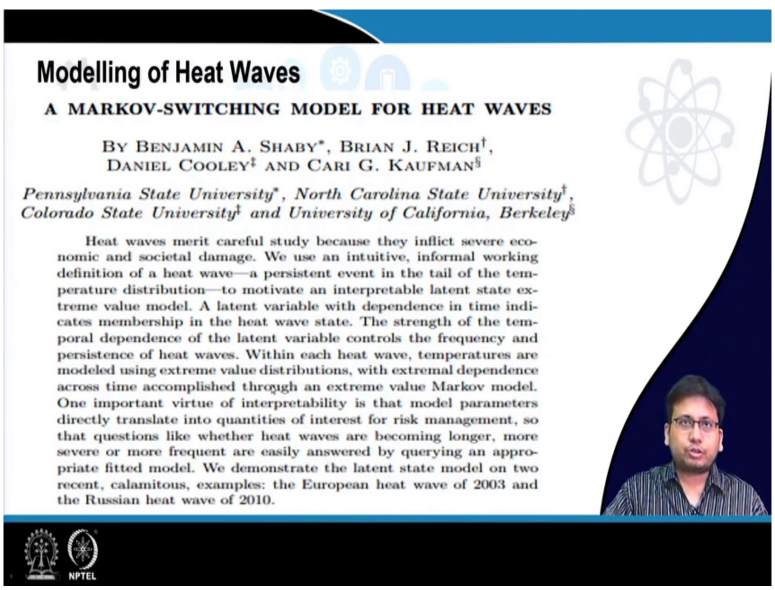
Now, these thresholds they can be either absolute or and fixed that is same for all locations or it can be location specific, that is, for some locations 40 degree Celsius might be the threshold for a some other locations 30 32 degree Celsius might be the threshold and so, on depending on whether its let us say its a hilly region or not I mean what is the standard in that region. That now these thresholds they should be flexible and they may vary spatially also.

What I mean by the threshold should be flexible I will come to that in a bit when we see a clear example. Also it need not be only about maximum temperature like humidity is also found to play an important role during heat waves that is the higher is the humidity the more intolerable a heat wave becomes and results in a higher mortality. So, and so, like earth scientists in recent times, they are also considering humidity as a possible factor while defining the heat waves.

The so, the approach here which we will discuss is to consider the weather state as a Markovian latent variable or $S(t)$. So, this is the latent variable because weather state by itself is not an observable variable, its not something that can be measured directly it is just a conceptual quantity which is its takes two value 1 or 0. If it is 1 it means that a heat wave is on and if it is 0 it means that there is no heat wave. And then there is the observation $Y(t)$ which depends on the observation of $Y(t)$ that is the daily maximum temperature which depends on $S(t)$ that is the current weather state as well as on $Y(t-1)$ that is the maximum temperature on the previous day.

The second it depends on the second thing on $Y(t-1)$ because like the maximum temperature is unlikely to change very drastically from one day to another like even if there like even if there is a heat wave or not.

(Refer Slide Time: 03:46)



So, this is the like. So, it is really the Markovian dynamics that is like that is in this paper that is which we are going to discuss the Markovian dynamics is inserted on this $S(t)$ variable. But the $Y(t)$ is also considered as a random variable which is conditionally dependent on $S(t)$. So, let us see how they model the whole thing.

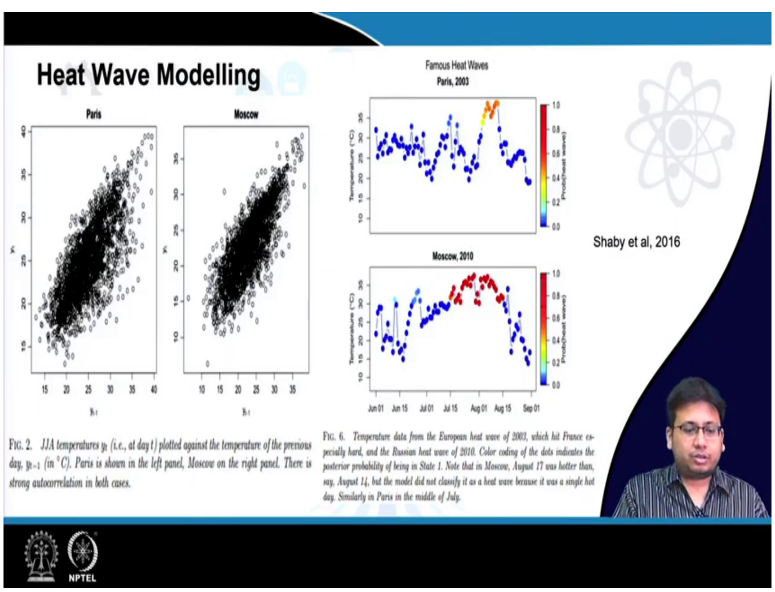
The heat waves merit careful study because they inflict severe economic and societal damage. We use an intuitive and informal working definition of a heat wave a persistent event in the tail of the temperature distribution; to motivate an interpretable latent state extreme event model.

A latent variable with dependence in time indicates membership in the heat wave state. The that is to say if $S(t) = 1$ then we will say that it is a part of a heat wave and otherwise not. The strength of the temporal dependence of the latent variable controls the frequency and the persistence of the heat waves. By persistence I mean the like for how many days at a stretch it remains active.

Within each heat wave temperatures are modeled using extreme value distributions with extremal dependence across time accomplished through an extreme value Markov model. One important virtue of interpretability is that model parameters directly translate into quantities of interest for risk management so, that will questions like, whether heat waves are becoming longer, more severe or more frequent are easily answered by querying an appropriate fitted model?

We demonstrate the latent state model on two recent calamities example the European heat wave in 2003 and the Russian heat wave in 2010.

(Refer Slide Time: 05:44)



So, before going into the model let us take a look at the data. So, like this is the maximum temperature in JJA means June, July, August the daily maximum temperature at Paris and Moscow.

So, here like basically we this is to study the relation between $Y(t)$ and $Y(t-1)$ that is like we earlier mentioned here that the observation $Y(t)$ depends on $S(t)$ and $Y(t-1)$ that is the previous days maximum temperature. So, here we see like as you can understand that we see a very linear relation between $Y(t)$ and $Y(t-1)$; that means, that is every on any given day the maximum temperature is roughly the same as the previous days maximum temperature that is why we cannot consider $Y(t)$ and $Y(t-1)$ as independent, we must consider some kind of a relation between $Y(t)$ and $Y(t-1)$ whenever we are developing a model.

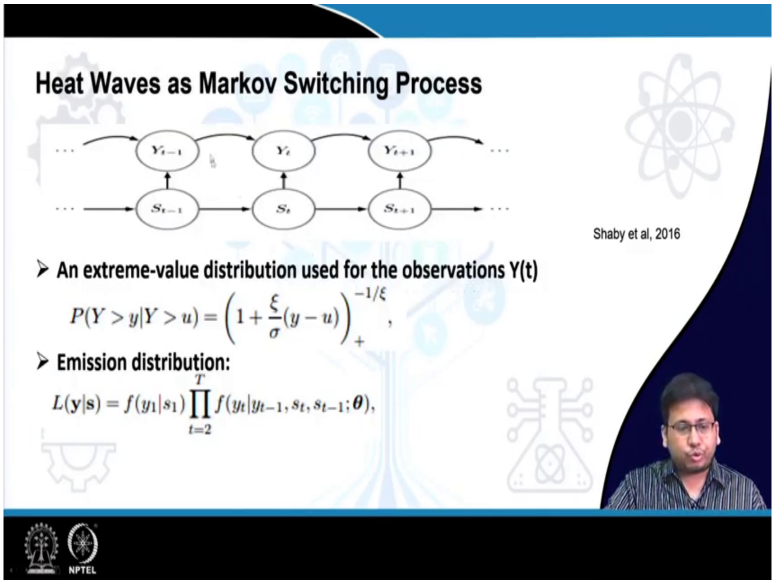
That independence is I mean that relation has to be strong enough than what can be captured if we do it like a hidden Markov model where $Y(t)$ depends only on $S(t)$. So, these kind of linear relation between $Y(t)$ and $Y(t-1)$ we see in case of both Paris and Moscow and now like these are the two use cases of the heat waves which they have considered.

Now, the interesting thing to for the heat these heat waves might be that see this is the heat wave and this is like one heat wave this is the other heat wave. So, as you can see if we if set some kind of a threshold like we it may not look like a its like a single heat wave. So, like even during this event there are not all the days you can see are equally hot there are some days there is a there is a period of three days where the temperature has actually slightly dropped.

But it may. So, if we define something like a threshold, the it might end up some of the intermediate days may end up dropping below the threshold, but that does not mean that these days are not part of the heat wave, they still are even if the temperature might be might have temporarily dipped a little bit the similar effect is seen in this case also.

In fact, here it is even more prominent like if you see like these few days they are also having quite a like similar temperature as the this heat wave period, but they are not I mean the they are not necessarily going to be considered as a heat wave.

(Refer Slide Time: 08:31)



So, now if we consider the Markov switching process which they have considered. So, these S these are the latent state variable which indicates the presence or absence of the heat wave and we see the corresponding observations of the daily maximum temperature.

So, the Markov we are calling it as a Markov process because $S(t)$ is like a or $S(t)$ or $Y(t)$ for that matter both are going to depend on only the previous observation not on any anything further. So, like we can say that $S(t)$ is independent of $S(t-1)$, $S(t-2)$ etcetera based on $S(t-1)$ that is if we know $S(t-1)$ then like that is enough to or that itself carries enough information about $S(t)$; considering past values we will not add anything.

Now, also note the difference between this model and hidden Markov model which we had discussed earlier. In case of hidden Markov model there is no relation between the observations $Y(t)$ is just a just depending on $S(t)$. $Y(t-1)$ depends just on $S(t-1)$ and so, on its S that the that does not mean that in a hidden Markov model $Y(t-1)$ and $Y(t)$ are independent the there is dependence, but that is through the S.

But in this case that the relation between a $Y(t)$ and $Y(t-1)$ that is quite strong and such a relation is unlikely to be captured through like indirectly through the S that too especially when S is binary if S has if S was itself a real valued variable then might be it may have still been possible to capture, but given that S is a like it is a binary variable which simply takes two values 1 or 0; that means, the variance of the emissions in each case is will be quite large and it will not be able to reflect the relation between $Y(t)$ and $Y(t-1)$ that we are that we observe.

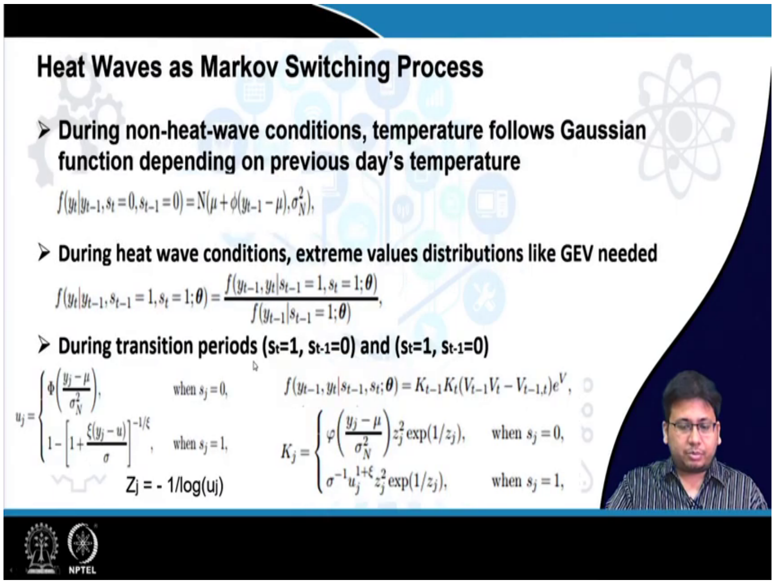
So, the extreme value distribution is used for this $Y(t)$ because these as you may remember these are the daily maximum temperature. So, like in an earlier lecture in module 1 where we talked about extreme value statistics we talked about extreme in two senses one is the block maxima and another was the peak over threshold. In case of peak over threshold we are talking about like that is about percentiles and so on, but in case of block maxima we are like there is a set of observations out of which we are considering the maximum values.

So, in this case also since we are talking about daily maximum temperature. Now in our day there are like we have many observations of temperature may be hourly and so, on even if it is hourly there are 24 observations, if it is per minute then we have like many more observations and that is we have 1440 observations per day etcetera.

So, like in each case that is we are talking about the maximum temperature. So, we have taking the maximum over a set of values. So, that is why in this case we should go for the block maxima approach and we have also discussed earlier that the block maxima that further follows what is known as the GEV distribution. So, like the also the GEV distribution.

So, the so, in case of GEV distribution that the like this is how it is defined that given that Y is high enough then the $P(Y > y)$ it follows this kind of a distribution. So, like where we have this the we like we have the different parameters like ε, σ etcetera. So, this is the emission distribution.

(Refer Slide Time: 12:31)



Then we so, when we are talking about $Y(t)$ ok. So, first of all how does this thing whole thing factorize that is. So, that is we have the time series of the Y variables as well as the sequence of the S variables. So, now, if I want to like express the sequence of observations Y as a function of the sequence of states S then this is how the whole thing factorizes.

So, both Y and S are random variables, but assuming that S is known this is how the distribution of Y it factorizes that is $Y(t)$ is basically its like it should it has to depend on $Y(t-1)$, $S(t)$ as well as $S(t-1)$ and so, now, we can say that in different situations we will have different distributions on Y.

So, suppose there is no heat wave in that if it is not a heat wave, then the daily maximum temperature can be simply we considered to be a Gaussian distribution. Why because like if there is no heat wave then we may not be able to say this kind of a situation that $Y > u$ that is that kind of threshold thing may not be there in this case.

But in like if in if the, but in case the there is no heat wave we may just consider it to be a Gaussian distributed variable, which of course, includes $Y(t-1)$ and so, on and μ is a like the parameter of our of the Gaussian distribution, σ is the variance parameter of the Gaussian distribution and so on.

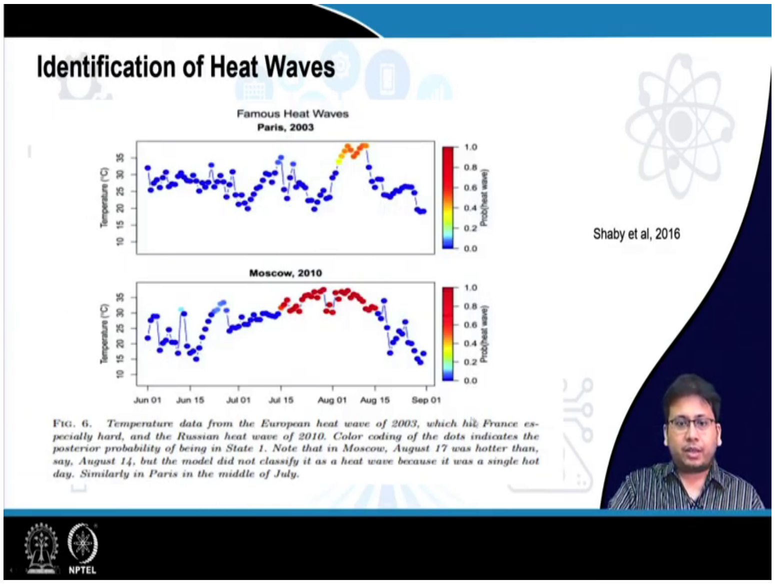
Now, during the heat wave conditions again that is the distribution of $Y(t)$ given its $Y(t - 1)$, $S(t - 1)$ and $S(t)$. So, $S(t - 1) = 1$ and $S(t) = 1$ means that it is the that in that t is the day t is inside a heat wave that is a heat wave has already started in that case like we like we can express it as this kind of a joint distribution.

So, like. So, we will see how to parameterize these kinds of joint distributions. And apart from that there are these transition periods where either $S(t) = 1$ or $S(t - 1) = 0$ means, a heat wave is just starting at time t or the reverse sorry this is the typo it will be 0 and 1 which indicates that the heat wave is ending.

So, in that case the like we use a surrogate variable called u or rather sorry not surrogate you can call it as a link variable called u. So, when $s_j = 0$ that is the heat wave has not yet started or has ended we will still use the something like the Gaussian distribution for the observation and once the heat wave has started or has not yet ended we will use the other distribution for the like for the observation.

So, this u this is transformed to that is the by this inverse logarithmic transition and like then they are like these kinds of mathematical transformations are carried out. So, these basically ensure that the variables I mean the observation variables, they follow the distributions which they are supposed to follow I mean the these kinds of marginal distributions they follow; the marginal distributions which we have considered in the both the heat wave a situation and the non heat wave situation and this is what the joint distribution looks like.

(Refer Slide Time: 16:14)



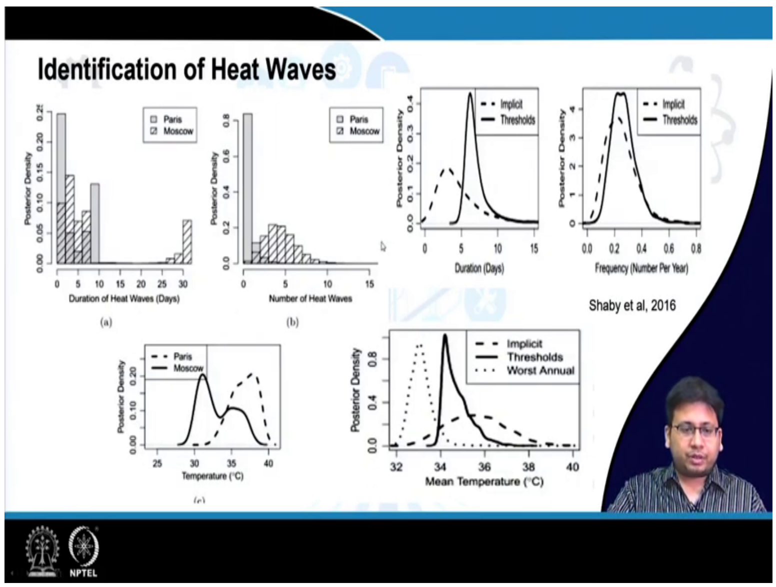
So, finally, what happens is that? So, like we have all these parameters. So, we need to estimate the values of the different parameters. So, like for the non heat wave conditions we have the μ and σ parameters of the Gaussian distribution for the heat wave situation we have the these σ and these ξ parameters. So, all these parameters need to be estimated.

So, what they do is they like they have observations. So, they use some like variable the some parameter estimation techniques such as the EM algorithm the expectation maximization and by which the which enables them to estimate the parameters in the presence of the latent variables S and once they have done it then they are able to like estimate the probabilities at the different at the different time points and so, as you can see the these color coding this red color this indicates the probability of having a heat wave on the different days.

And as you can see in these days in the middle where the temperature has actually dropped to normal levels or to near normal levels even though they are within the heat waves as you can see they are still shown to be to have high probability as being part of the heat wave. Why is that the case that is because we are still considering the previous days temperature $Y(t-1)$ and because the in the previous day the heat wave was still active.

So, although on this particular day the temperature may have fallen because of the previous days influence, we are still like assuming that is we can still guess that the heat wave is still active it has not ended yet. So, and similarly then like if you consider these days its possible that they are also part of a heat wave, but with a much lower probability. Similar thing we see in the Paris heat wave also.

(Refer Slide Time: 18:21)

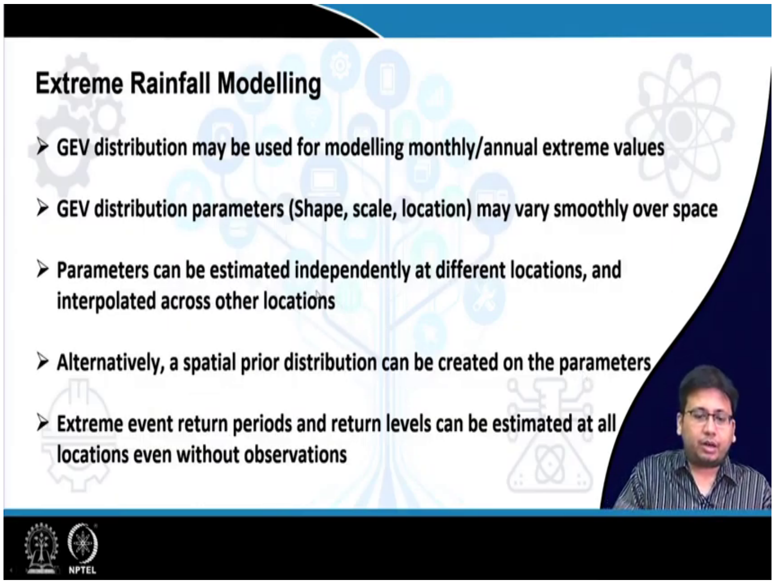


And then once we have like that is when once we apply these technique to identify the heat waves, we can also study the different properties of the heat wave and compare them with like what is actually observed. So, here they have like the different properties of the heat wave they have considered as the durations of the heat waves.

So, in different locations a heat wave may typically have different duration some locations may have short heat waves some other locations may have long heat waves. So, they have study studied that the distribution of duration in different locations according to their model as well as according to and they have they can compare the results obtained from the model to what is actually observed.

Then apart from that other parameters are number of heat waves that maximum temperature during the heat waves and so on and so forth.

(Refer Slide Time: 19:18)



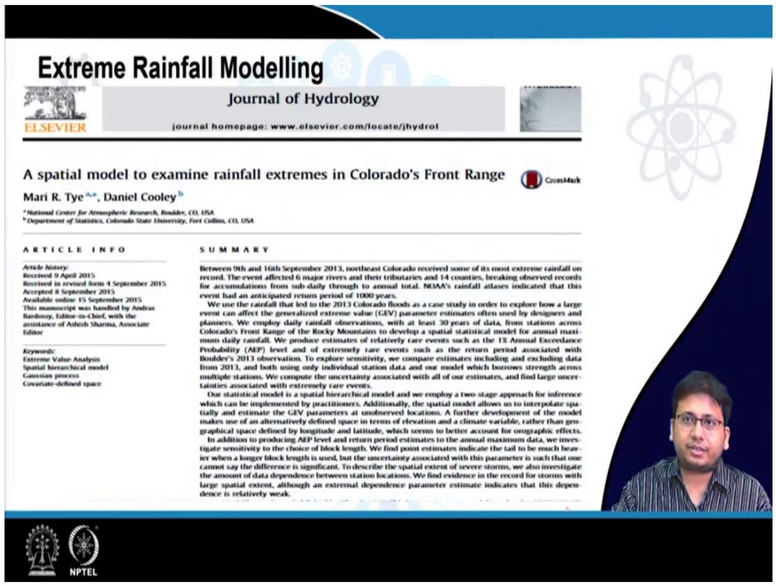
Now, apart from heat waves. So, this extreme value modelling this can also be used for extreme rainfall modelling. So, the GEV distribution which we talked about is like for the block maxima case, the GEV distribution may be used for modelling the monthly or annual extreme values and these the GEV distribution parameters if you remember the GEV distribution, it has like three parameters the shape, scale and location and these parameters they may vary smoothly over space.

So, the parameters can be; if we some let us say we have observations at some locations which might be some isolated locations, may be different cities of a country and so on. In the intermediate places we may not have observations. So, from whatever places, we where we do have observations let us say in the four cities, then in each of those four cities we can estimate the parameters independently and then use some kind of interpolation to estimate the parameters at the other at any arbitrary location in that region under the assumption that the parameters vary smoothly over space.

An alternative approach is to define some kind of a spatial prior distribution and that is instead of measuring them independently at the different locations let us say that there is a prior distributions something like a Gaussian process and the parameters at all locations including those locations where observations are available, the these parameters are actually drawn from that kind of a that spatial distribution only.

And then this approach allows us to estimate extreme event return periods and return levels at all locations even at those locations where there are no observations.

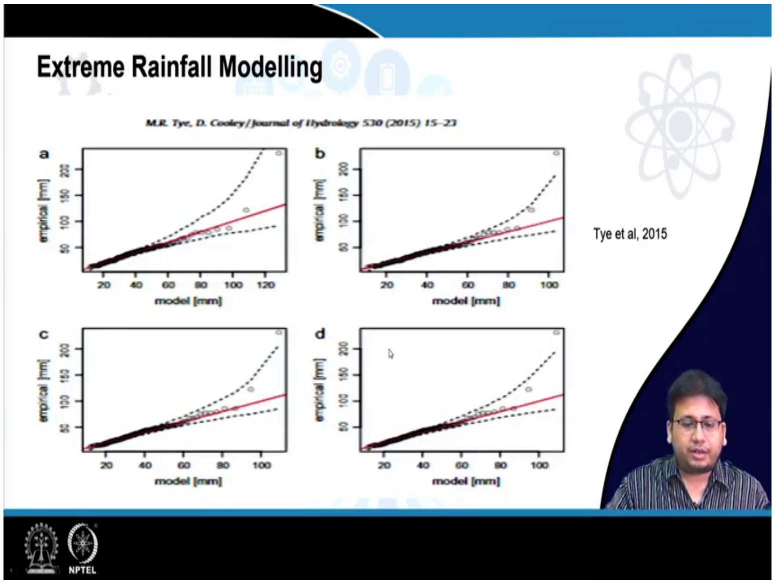
(Refer Slide Time: 21:17)



That is at any given time I may not know whether an extreme event is happening there or not because I cannot observe, but in a statistical sense I can know how extreme an extreme event is likely to be there that is like when we talk about return level I am just ask I am basically asking like how much is a 100 year flood or how much is a 100 year heat wave in that location.

That is say for or if I may frame it in a differently that is what is the one the 99^{th} percentile flood or what is the 99^{th} percentile heat waves that might happen in any given location? So, the first approach which we talked about that is measuring the GEV parameters independently and then using spatial interpolation was used in this particular paper like. So, this is by Professor Daniel Cooley he is one of the most active researchers on this spatio temporal extreme events.

(Refer Slide Time: 22:24)

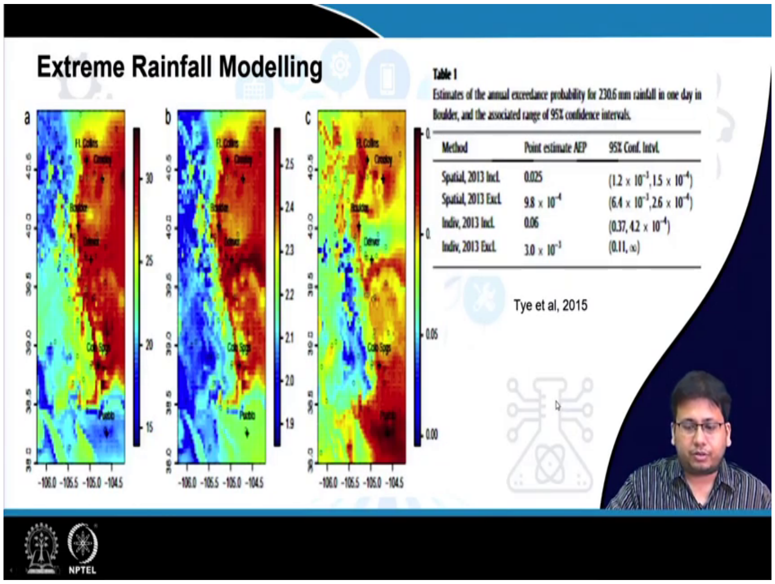


So, here they use this GEV distributions in a particular region and they actually like carried out the parameter estimations of the GEV distributions and then used something like interpolation to like for all the other for other regions also where they do not have observations. So, this is what the model fit look like where they like for that is they are.

So,. So, if you remember the QQ plots which we discussed when we are talking about extreme. So, this is something like that. So, like here they like they are basically comparing the QQ plots of the observations as well as the models and. So, ideally the QQ plot should go like this according to this red line.

So, in different models which we are seeing that in some cases the model slightly underestimates the observed extremes and in like in a different model it may slightly over estimate the, but the like. So, if we consider some kind of an ensemble of these models we will we are more likely to get closer to what is observed.

(Refer Slide Time: 23:23)

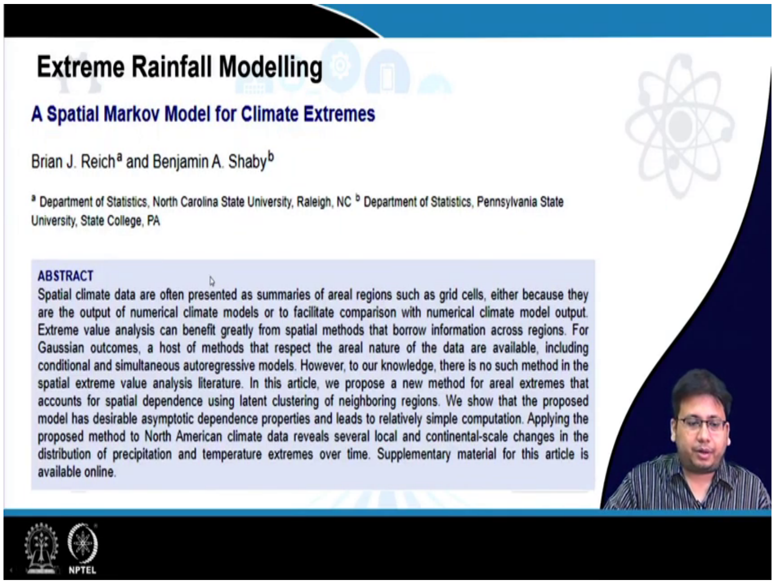


So, this is what the like their interpolation may look like. So, here like you have you can see some plus signs. So, those are the they are the different cities like Fort Collins Boulders etcetera. So, these are all cities in the Colorado state of US.

So, at these locations they have sensing facilities where they can actually measure the amount of rainfall in other places they may not be able to measure it, but they are still estimating it because they have like by using the approach of interpolation. And also when they are doing it they have to take into account other covariates that is when they are estimating the parameters at some at other locations also like as you can see so, there are these red regions and the blue regions.

So, the these blue regions are like on the mountains that is they are at a higher altitude. So, at these higher altitude the extreme rainfall parameters can be somewhat different because of the because the altitude itself plays a role. So, they actually like take that altitude as a factor when they are estimating the different spatial parameters.

(Refer Slide Time: 24:34)



So, and now another paper on this same topic. So, like as I said earlier rather than estimating the parameters independently at the different locations, we can also do it by like considering the parameters to be part of some kind of a prior spatial distribution. So, that is what is done in this paper.

So, here what they are saying is spatial climate data are often presented as summaries of areal regions such as grid cells either because they are the output of numerical climate models or to facilitate comparison with numerical model the climate model output, that is, when we like when some kind of a process model is run, it is usually run on a grid structure that is there are no point wise observations and, but, but actual observations are always point wise they are never gridded.

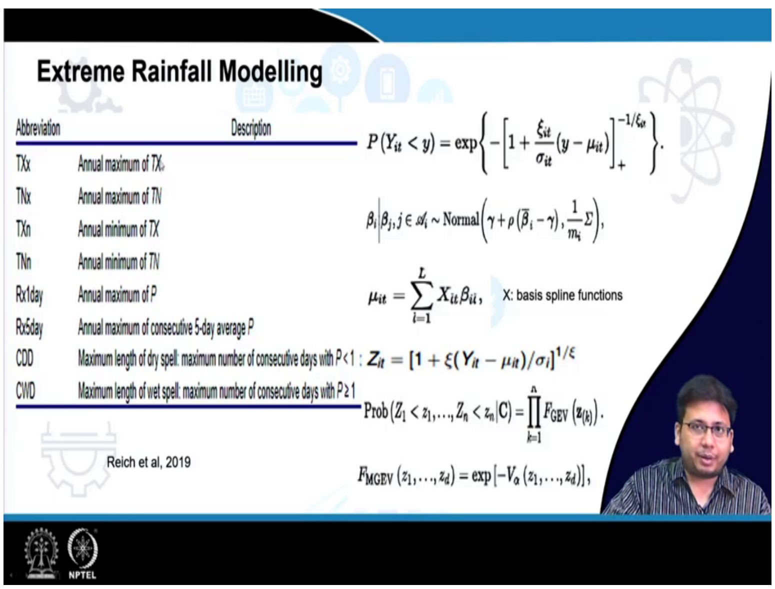
So, now, these gridded observations they are basically summaries we can say over a like over a particular region. So, we may have measurement several measurements in a region and the gridded representation is something like a some the mean of those observations and so, on.

Now, extreme value analysis can benefit greatly from spatial methods that borrow information across regions. For Gaussian outcomes a host of methods that respect the areal nature of the data are available including conditional and simultaneous autoregressive models. However, to our knowledge, there is no such method in the like in the spatial extreme value analysis literature in this article we propose a new method for areal extremes that accounts for spatial dependence using latent clustering of neighboring region.

By areal extremes what they mean is like the spatial block maxima earlier we like when we talked about block maxima earlier, we are talking about temporal block maxima now here its talk like this is talking more about spatial block maxima that is we have observations at different points and then we are and they may have their individual maxima, but how about the summary maxima of the entire areal region which is which we need for the gridded representation.

We show that the proposed model has desirable asymptotic dependence properties and leads to relatively simple computation. Applying the proposed method to North American climate data reveals several local and continental scale changes in the distribution of precipitation and temperature extremes over time.

(Refer Slide Time: 27:31)



So, like these are the different variables which they have considered the annual maximum temperature, annual minimum temperature, the annual maximum of TX.

So, TX means the daily maxima, TN means the daily minima. So, the TX_x means the annual maxima of daily maxima, TN_x means the annual maxima of daily minima and so on. Similarly we have the annual minima of the daily maxima and annual minima of the daily minima also similarly they have for this is for temperature similarly for precipitation they have the annual maxima, they have the maxima of the consecutive 5 day average, they have the maximum length of dry spell, wet spell etcetera.

So, at any given location like we they may have this kind of a GEV distribution which includes the GEV parameters like the location, shape and scale. So, as you can see it means i may mean a particular location and t may mean a particular time and also this μ_{it} the that is. So, so basically what they are going to do is they like for these different parameters they are trying to put something like a spatial structure on it.

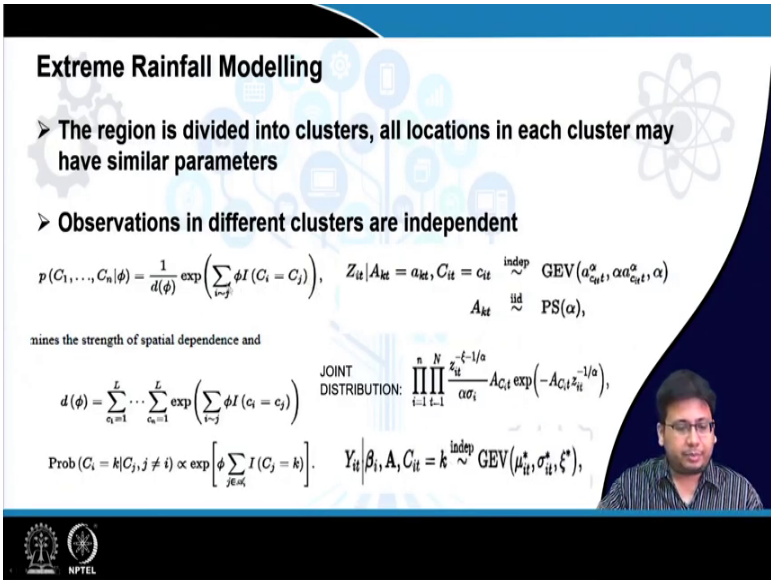
So, like if you consider this μ_{it} that is expressed in this way that is like they have some basis functions X_{it} and the μ at any given location and time is basically a linear combination of the these different basis functions. Now the coefficients of this linear combination they may vary from location to location which we need to estimate or the these basis spline functions we may like some standard values may might be used or if necessary we may want to estimate these the basis functions themselves from data which will of course, be more difficult.

So, now this so, we have the $Y's$. So, we can actually transform the Y to Z by doing some kind of transformation. So, that the distribution of Z_{it} that is. So, the so, basically this is the transformation which I talked about and then like we can have a this kind of a joint distribution in terms of the $Z's$ instead of the in terms of the Y. So, the like somewhat something like this and like of course, assuming the like.

So, there are n different locations. So, that the time for the time being let us drop or let us just assume that this is like either it is like completely time independent or like its constant over time. So, let us drop the t for the time being and focus only on the locations. So, at like there are n locations at which we have these measurements and so, we have the Z variables which we consider them to be independent and we write it into in the in this way.

So, they are, but they are the observations are independent, but the parameters are not. So, like what is necessary is to somehow like estimate the these parameters.

(Refer Slide Time: 30:47)



So, for that purpose what they do is they divide the region into similar clusters and all locations in each cluster may have the similar parameters that is its. So, basically they define something like a clustering problem and the cluster variables is defined in using these I mean the. So, they introduce this cluster variable $C_1, C_2, \ldots, C_n$.

So, each of the locations they are associated with a with this C variable which indicates which cluster it is a part of. So, these cluster variables they can take for any configuration any cluster configuration like this indicate this is the probability of it. So, like $C_1 = 1, C_2 = 1, C_3 = 2$ this is one configuration then $C_1 = 1, C_2 = 2, C_3 = 2$ this is another configuration.

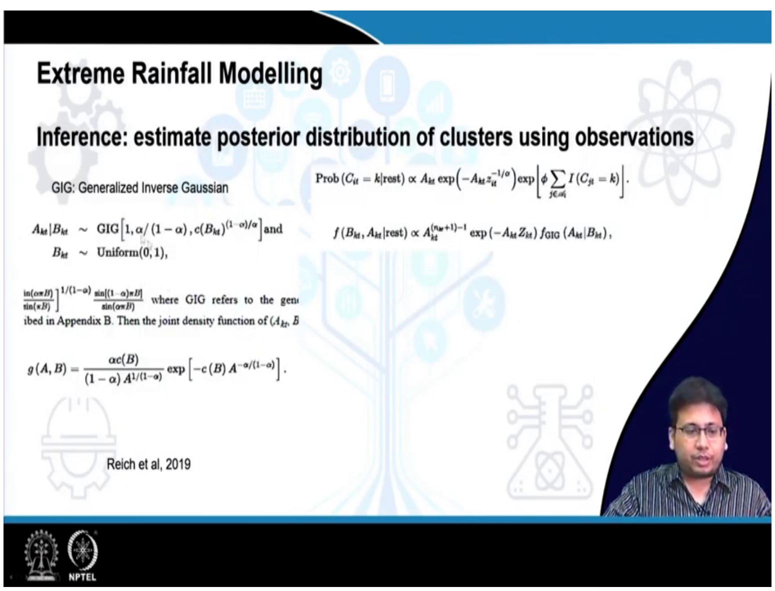
Then $C_1 = 1, C_2 = 1, C_3 = 1$ this is another configuration and so, on. So, each of these cluster configurations each of them will have some kind of a probability and that is specified by what is known as the Potts model here and then like. So, based on that they can determine the strength of the spatial dependence also like using this $d(\Phi)$.

So, like basically what they need to do is, they needful need some kind of a conditional dependence because we do not know these all the cluster variables. So, they like we need to estimate each of them basically we need to do something like a clustering. So, what they do is, the clustering is done not by a deterministic approach like k-means, but by a probabilistic method. So, for every variable that is we try to estimate its like its cluster coefficient conditioned on the cluster assignments of the like other locations and that is done using like this so, for that we have a probability distribution like this.

So, we can say that this is something like a prior distribution on the cluster of every location. So, then we so, basically when we have the observation Z_{it}. So, like we can a like a it follows the GEV distribution according to the parameters like this where the parameters are like they follow like each of the parameters are like they are like they are dependent provided they are all in the same cluster that is like all let us say the c^{th} cluster for.

So, for all the locations within the cluster the this Z value they follow one particular distribution. So, they have the so, we can come up with this kind of a joint distribution and it can be shown that if this is the case in that case the actual distribution Y_i which is actually what we observe the maximum values that we observe before all these transformations and so, on it follows a GEV distribution what it is expected to follow?

(Refer Slide Time: 33:54)

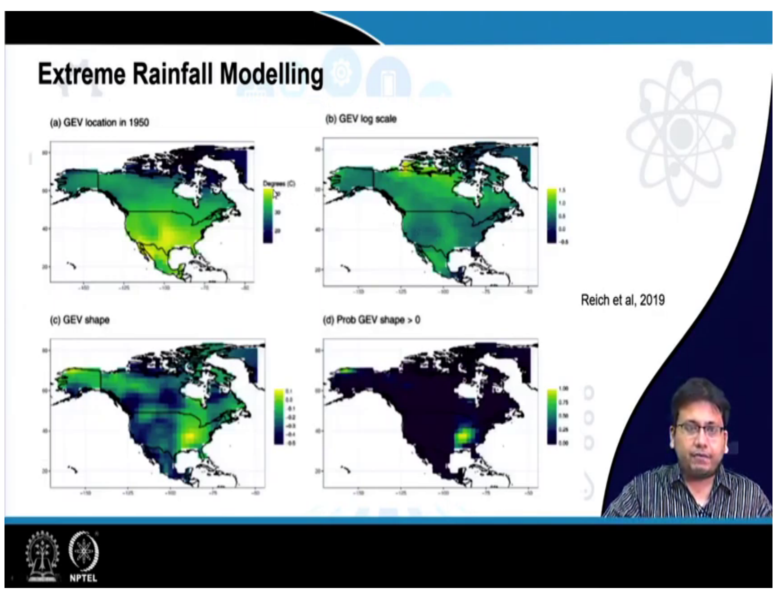


So, then what is required is like we need of course, need what we need is now to estimate the parameters as well as the latent variables like C, the clusters. So, for that they we have to create something like a posterior distribution on the clusters.

So, like earlier we had a prior distribution on the on the cluster variable, now we have a posterior observation on each of the clustering variable based on not only the other clusters, but also the observations Y or Z whatever you call it. So, this is the that posterior distribution and so, like and apart from that for these transformation matrices B, A etcetera like which were like used to define these the parameters of the GEV in a any particular location.

So, these also have to be estimated based on the data and the all the clusters. So, we do something like Gibbs sampling where these different sets of random variables. So, A, B, C these are the latent random variables and the then those α, σ etcetera those are the different parameters. So, we do the Gibbs sampling here to that is in the way which we have discussed in another lecture to estimate all those.

(Refer Slide Time: 35:14)



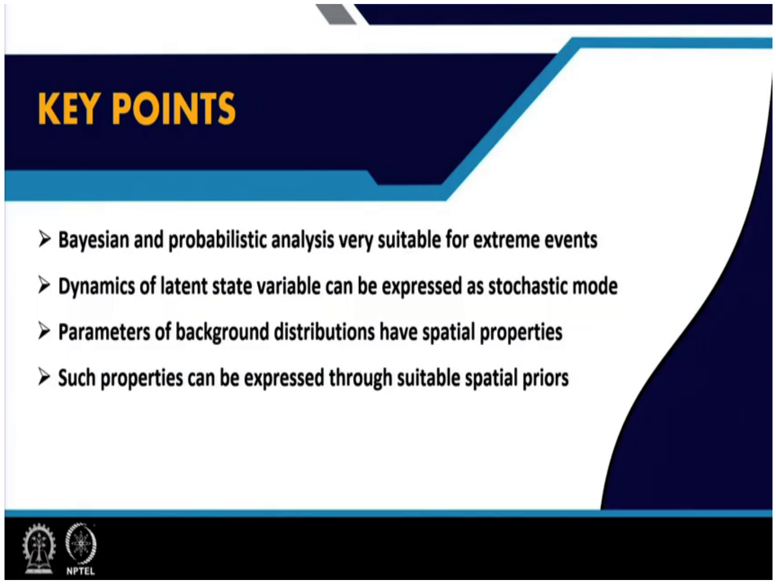
And once we have all the parameters then we can like at any given location we can like that is like we can plot the location, shape, scale etcetera parameters of the GEV distribution. So, here they have created a map of the different GEV parameters like this.

So, like for the scale they have this is the map of the scaled parameters of the GEV, you can see a reasonably smooth variation over space. This is the map for the location parameter of the GEV again you can see some smooth variation of what shape and so on. It will be interesting to see what are the locations which sorry what are the regional clusters which they have obtained.

So, for example, this might be one regional cluster, this might be another regional cluster, this might be a third regional cluster and so, so on and so, forth.

(Refer Slide Time: 36:08)

(Refer Slide Time: 36:11)



So, these are the references of the papers which we discussed today. So, the key points to be taken home from this lecture are first of all the Bayesian probabilistic and probabilistic analysis are very suitable for understanding these extreme events.

The dynamics of the latent state variable can be expressed as stochastic model and the parameters of these background distributions they themselves have some spatial properties which can be expressed through suitable spatial priors as was discussed like here like this in by this kind of a like regionalization analysis. So, that brings us to the end of this lecture we will in the following lectures we will see a few more use cases of machine learning to on earth more insights of the earth system processes. So, till then bye.