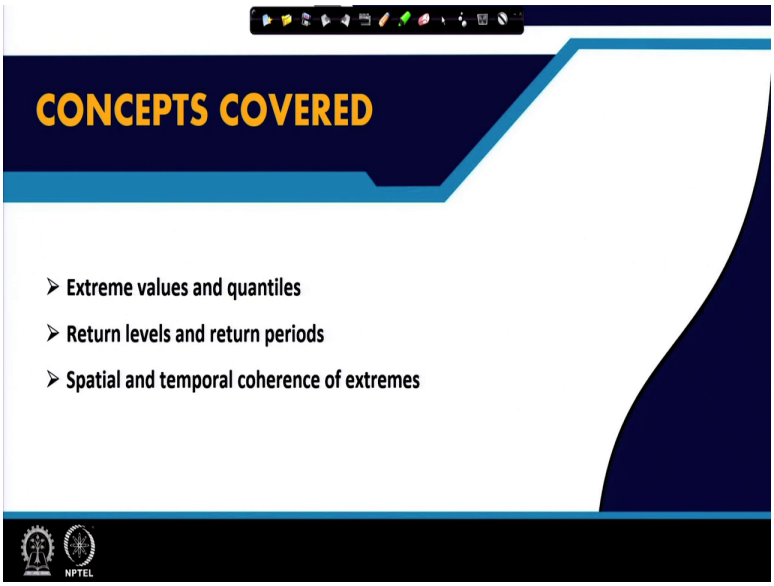
Machine Learning for Earth System Sciences
Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 01
Spatio - Temporal Statistics
Lecture - 06
Extreme Events

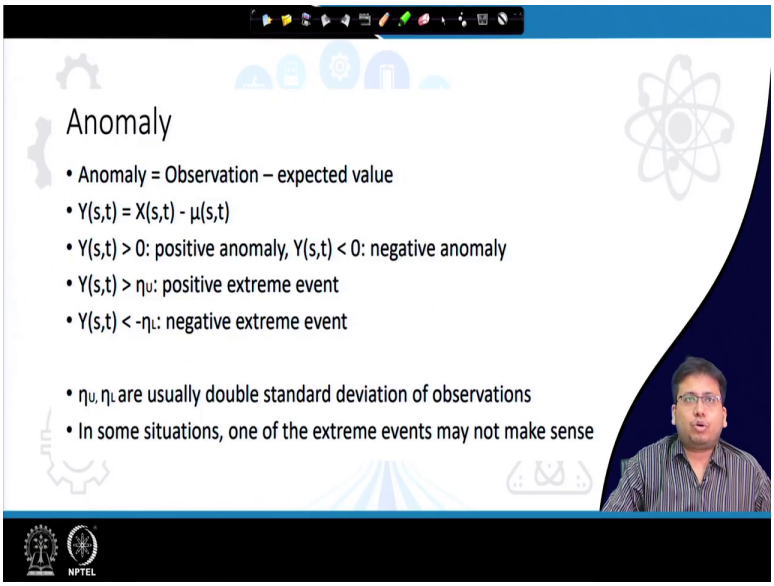
Hello everyone, welcome to lecture 6 of this course on Machine Learning for Earth System Science. We are continuing the module 1 which is on Spatio Temporal Statistics. So, today we are going to talk about Extreme Events.

(Refer Slide Time: 00:44)



So, like you can consider this as an extension of the previous lecture which was on anomaly events. So, the concepts we are going to cover today are Extreme values and quantiles, Return levels and return periods and Spatial and temporal coherence of extreme events.

(Refer Slide Time: 01:01)



So, let us begin with the definition of Anomaly, which we already covered in the last lecture. So, anomaly is basically just the deviation of the observe of an observation from it is long term expected value. So, suppose the observation is at any given location and time is $X(s,t)$ and it is climatology is $\mu(s,t)$like we had discussed in the last lecture that this might be the spatial or temporal climatology also.

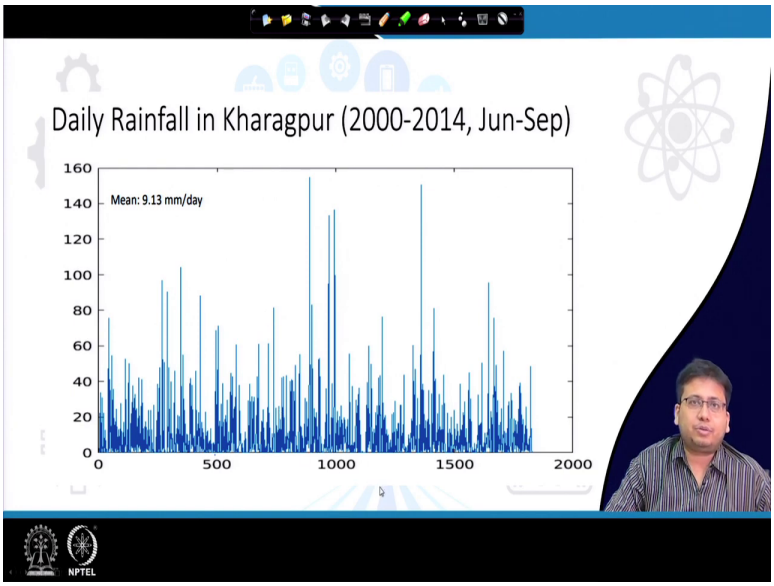
Now, if so by subtracting them what we get is $Y(s,t)$ now if , $Y(s,t) > 0$, then we call it as a positive anomaly and if $Y(s,t) < 0$ we call it as a negative anomaly. However, if the anomaly is above a particular threshold let us call it as η_U, then we may call it as a positive extreme positive extreme. Similarly if it is a negative anomaly and it is below some particular negative threshold, then we can call it as a negative extreme.

Now, what these threshold are η_U, η_L etcetera that is a like it is again it is subjective and context dependent, like we like it is sometimes people use the double standard deviation as a measure of these η_U and η_L. But that is not strictly necessary and it is possible that in some of the situations that one of the extremes may not make any sense. Like for example, in case of daily rainfall.

So, daily rainfall like positive extreme makes sense like if you are if we have a very large quantity of rainfall we call it as like as like if there is 100 millimeters of rainfall at a given location on a given day, we can call it as a positive extreme. But there is no negative extreme to be defined in this case I mean rainfall can never be negative and if it is 0 rainfall that simply means a day in which it did not rain there is nothing extreme associated with it.

So, in that case only positive anomaly makes sense, but if we are considering let us say annual rainfall then both positive and the negative anomalies may make sense.

(Refer Slide Time: 03:14)

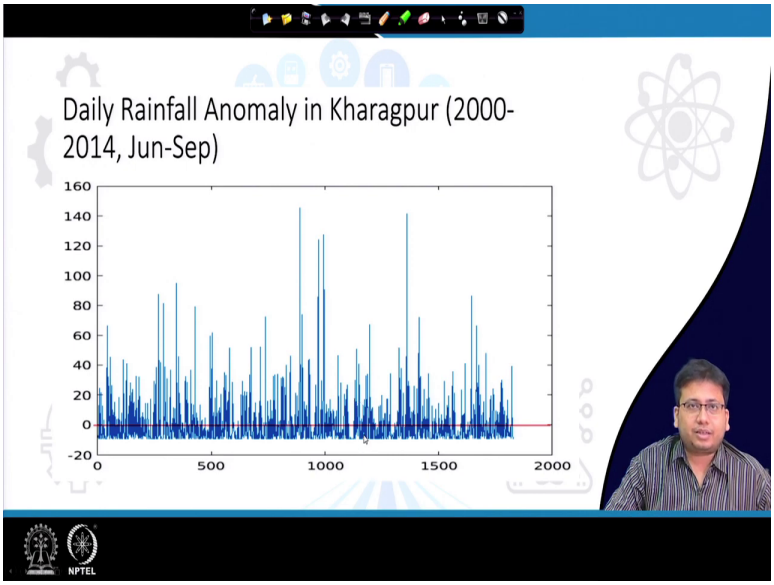


Now, to understand that let us look at the daily rainfall time series of Kharagpur for in the months of June to September for this period of 15 years from 2000 to 2014. So, the total number of days here is. So, there are 15 years in each year from 1st June to 30th September there are 122 days. So, the total number of days we are dealing with here is $15 \times 122 = 1830$, that many data points we have and at each of those days we have these measurements of rainfall. So, why did we choose this particular period June to September?

That is because this is the monsoon season and most places in India including Kharagpur receive at least 80% of their annual rainfall during this monsoon period only ok. Now the mean rainfall here the daily mean is 9.13 millimeters per day. So, as you can understand that on some days it is

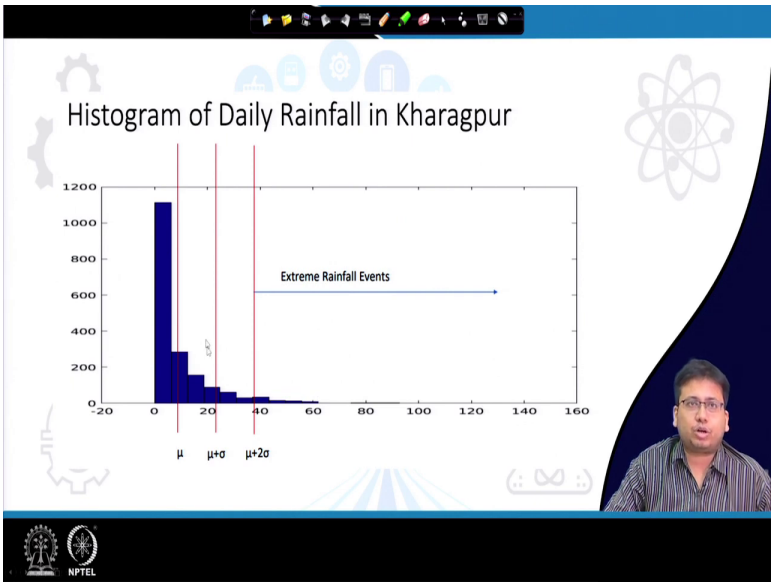
the rainfall is above that mean and some which means the positive anomaly and on some other days it is below that mean which is the negative anomaly.

(Refer Slide Time: 04:27)



So, now if we like put the threshold at 9.13. So, like these days which are which have these are the days which have the positive anomalies and these are the days which have the negative anomalies.

(Refer Slide Time: 04:43)



Now, let us say we plot the histogram of this of these values ok. So, like you can understand that on most of the days the values are actually quite low. So, that is why it is 0 or close to that.

So, we see that the it is the small values which have the maximum frequency and higher the values are the lower is the frequencies ok. In fact, like it is quite rare for any day to have let us say more than say 40 millimeters of rainfall. Now on this kind of a histogram let us put the value of μ the climatology. So, that as we already said that value is 9.13.

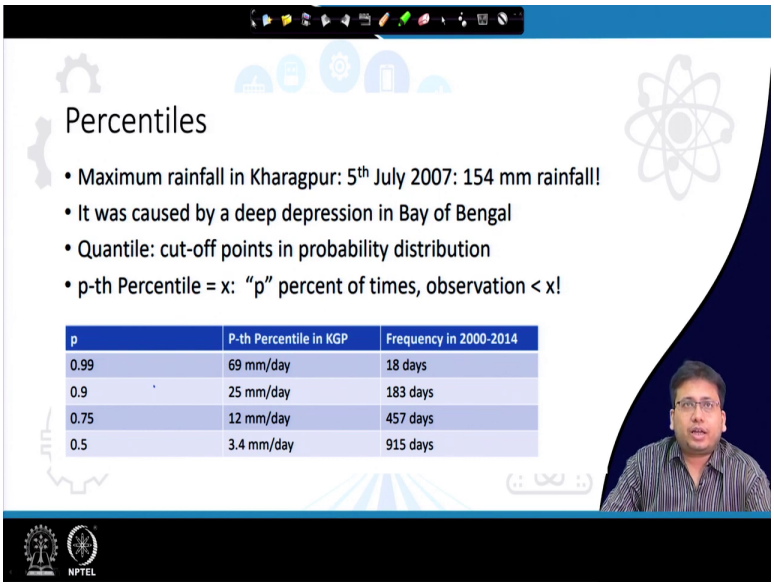
So, here like we have drawn this red vertical bar to indicate that these are the like negative anomaly incident events and these are the positive anomaly events. Now apart from mean let us also calculate the standard deviation σ. So, like let us say that we have some value of standard deviation about say 12 or something like that. So, or may be slightly higher may be 15.

So, $\mu + \sigma$ is like is the value somewhere here. So, here we have put another red bar and similarly $\mu + 2\sigma$ that is like we have put another red bar here. So, like we can say that all values that are beyond any of these thresholds we can consider them as the extreme events. So, like the standard practice is to choose $\mu + 2\sigma$ as the cut off or the threshold, the η_U as we are saying for the extreme positive extremes.

So, like every day in which the rainfall has been above this particular threshold which is say something like 40 millimeters, all those days will be considered as extreme rainfall days ok so, but you can see from this distribution how rare those extreme rainfall events are. Most of the days are actually below the that is they have negative anomaly on like only a small number of days actually have a positive anomaly.

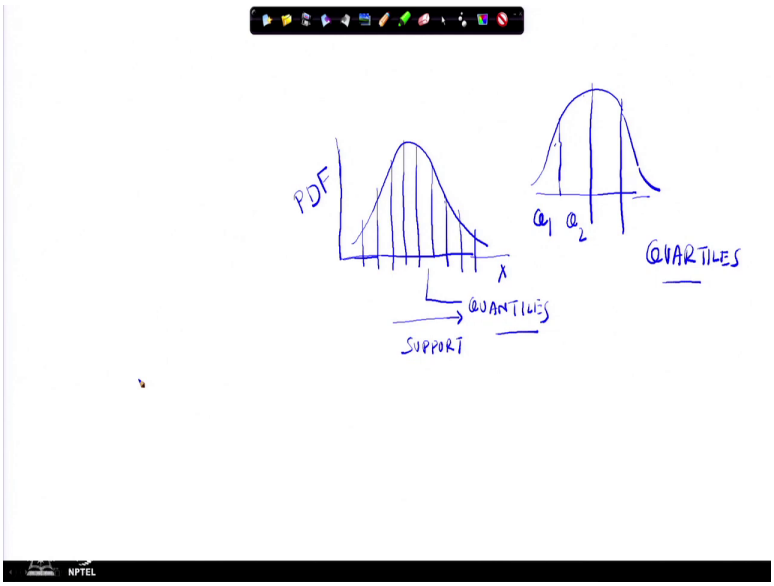
And then among those positive anomaly days a very minuscule fraction of the days actually have extreme rainfall events, if extreme is to be defined in this way. If extreme is defined in this way as $\mu + 1\sigma$ then the number of anomalies I mean extremes will of course, be slightly higher. So, similarly we can define some other thresholds also for extreme.

(Refer Slide Time: 07:36)



Now, that brings us to another concept for defining these extremes that of percentiles or quantiles. So, quantile is basically, so suppose we have in some kind of a probability distribution let us say something like a PDF. So, like if we draw like a figure like this.

(Refer Slide Time: 08:02)



Let us say this is some kind of a probability distribution ok. So, what we can do is we can like we can actually divide this range of values. So, like on the $X - axis$ what we have is basically the support set. Now this and on the $Y - axis$ we are having the PDF, now we may choose to divide this support set into certain sectors like this those sectors may be like at like any regular intervals. So, each of these sectors these are known as the quantiles. Now there are certain special ways of defining quantiles. So for example, there are quartiles like I may divide the whole thing into 4 like 4 parts.

Let us say this is the first quarter then this is the second quarter the third quarter and the fourth quarter. So, this can be called as the first quartile, the second quartile etcetera. So, these are the quartile, so note that quartile is different from quantiles. Similarly, we can also define percentiles. So, percentile is something which you are all familiar with in case of like because the results of these competitive exams like gate or IIT JEE etcetera are often expressed in terms of percentile.

So, percentile is when you divide the whole supports into a 1000 bins or a 1000 sectors like this and. So, like so that so like you can say that a percentile is a special case of quantiles just like quartile is a different case of this percentiles. Similarly there are deciles and things like that like this is related to the concept of median. So, median is like you have a range of values.

The median is the like a exactly that value such that 50% of the values are above it and the other 50% of the values are below it. So, similarly just like just extending the concept of medians we get the concept of quantiles right. So, like in this case again we use the idea of quantiles especially percentiles.

So, like let us say that let us consider the 99^{th} percentile; that means, 99% of the days will have lower value of rainfall than a particular value. So, in this case, the 99^{th} percentile or like instead of calling it as a 99^{th} percentile we can you can just call it as 0.99. So, like you can say that with 0.99 probability the rainfall on any given day is going to be less than a particular value. So, it if you do an analysis of these values which we considered here you may find that like on in this data set, if you do an analysis of the values you will find that cut off value is 69 millimeters.

There that is there are only 1% of the days in which the rainfall is more than this 69 millimeters in other words, like if you consider any random day any out of those 1830 days. If we choose any day at random there is only 0.1 probability that the rainfall in on that day will exceed 69 millimeters ok and it is actual frequency out of those 1830 days is 18 days. So, as you can understand one roughly 1 by 100.

So, similarly the 90^{th} percentile is 25 millimeters, 90^{th} percentile that roughly means that there is 0.9 probability that if you pick up any day at random the rainfall in it is going to be less than this threshold of 25 millimeters. And it turns out that there are actually 183 days in which the rainfall has exceeded this 25 millimeters; 183 out of 1830, so that is just 10%.

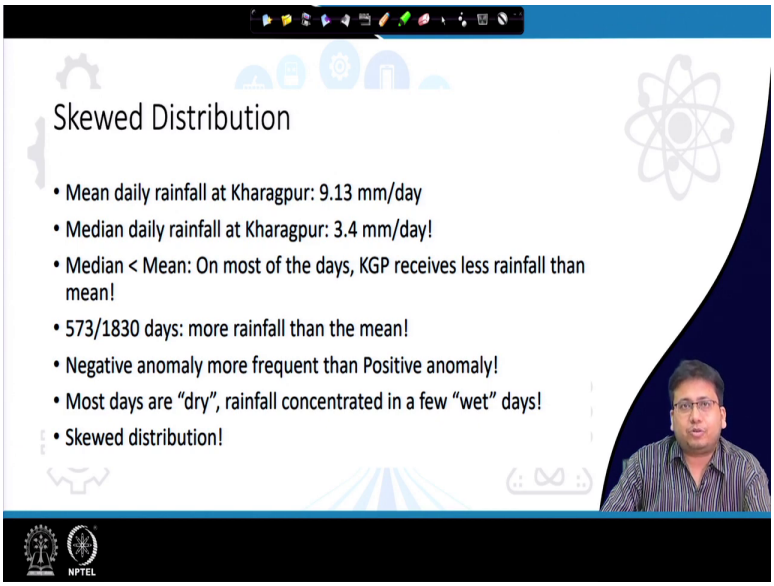
Similarly we can define the 0.75 or the 75^{th} percentile, the 50^{th} percentile and so on and 50^{th} percentile as you can understand is just the median. So, there are 915 days in which the rainfall is more than it and another 915 days in which the rainfall is less than it and it turns out that the median value is 3.4 millimeters.

The interesting thing to note here is that the median rainfall is 3.44 millimeters while the mean rainfall we had earlier seen is much larger, it is 9.13 millimeters. That means, what? That means, it is a very skewed distribution, in case of Gaussian distribution we know that the mean, median and mode all of them coincide. That is there are like what I mean or the is the it is like the expected value of the Gaussian distribution is also the most frequent value.

And like you can expect that the number of samples which will have lower values is roughly going to be equal to the samples which will have higher values, that is why Gaussian distribution is called the perfectly symmetric distribution. But in this case clearly that this distribution is highly asymmetric, you can see that the it is the low values dominate, but there are some very high values also like you can see some values like 80, 100 etcetera.

In fact, the maximum value is 154 millimeters. So, how high is that compared to the mean? So, it is like we can say that in most days or on at least 50% of the days the rainfall is actually below 3.4 millimeters, but there are some very rainy days also which have pulled the average from 3.4 to all the way till 9.13 ok.

(Refer Slide Time: 14:38)



So, this is our typical example of a like a long tailed distribution. So, there are probability distributions have very there are some probability distributions which are symmetric like the Gaussian and then there are some which are skewed or asymmetric like this one, I mean this histogram you can consider to be the PDF of some probability distribution.

So, you can understand that this is the like it is a skewed distribution, but it is also a long tailed distribution. What do I mean by that? That means, there are various tail events that is which are

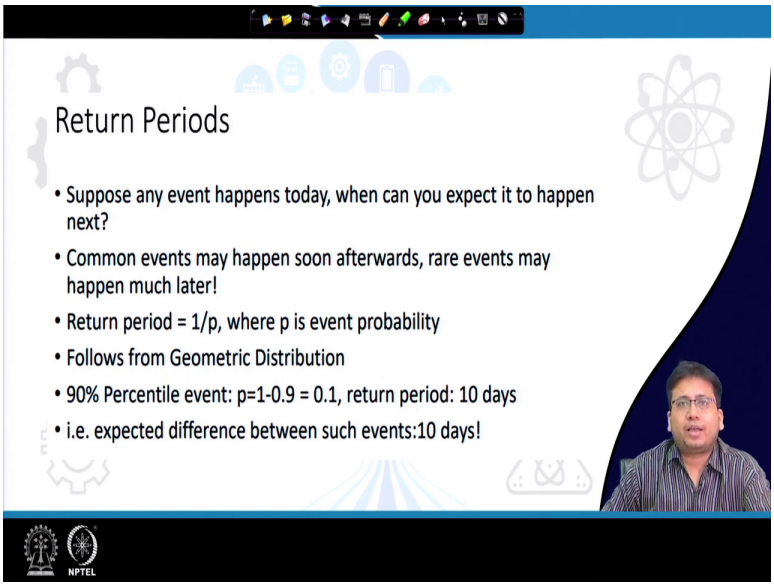
very far away from the where the bulk of the probability mass lies. But like say values like 154 and things like that they are very far away, but they still have some probability mass.

So, that is like called the a tail of the distribution. So, the like as you can see the this thickness of the histogram this vanishes at 40 millimeters, but it is not that higher values do not occur. In fact, a value as high as 154 millimeters also occurs, that is why we say that it is a distribution with a very long tail ok. Now many geophysical variable have this kind of a long tailed property.

Now, so like as I said that there are only 573 days in which the rainfall is above the mean that is they have some kind of positive anomaly while the remaining days, so, this 573 out of 1830 this will be roughly about 30% may be 31% or something like that and the remaining 69% of the days are having a negative anomaly.

That means the negative anomalies are much more frequent than the positive anomalies; in other words, that means, that most of the rain days are dry and most the bulk of the rainfall is concentrated in a few wet days ok.

(Refer Slide Time: 16:39)



So, that actually so like for those of us who have first-hand experience of the Indian monsoon we will understand that, that is actually like we can corroborate with our daily experience.

We know that even during monsoon season it is not that it is raining all the time. There are many days in which it does not rain. But there are some days in which it rains heavily and makes up for that and like in there is also the hypothesis that one of the impacts of climate change is going to be this effect is going to be more and more accentuated. That is we will see a larger fraction and even larger fraction of these dry days or negative anomalies.

And an even and but the days in which they will be positive anomaly the number of these extreme events are also going to increase. Now this brings us to a concept of Return Period. So, suppose an event happens today, suppose there something very strange happens today let us say there is suddenly 100 millimeters of rainfall. So, you can understand how rare that is.

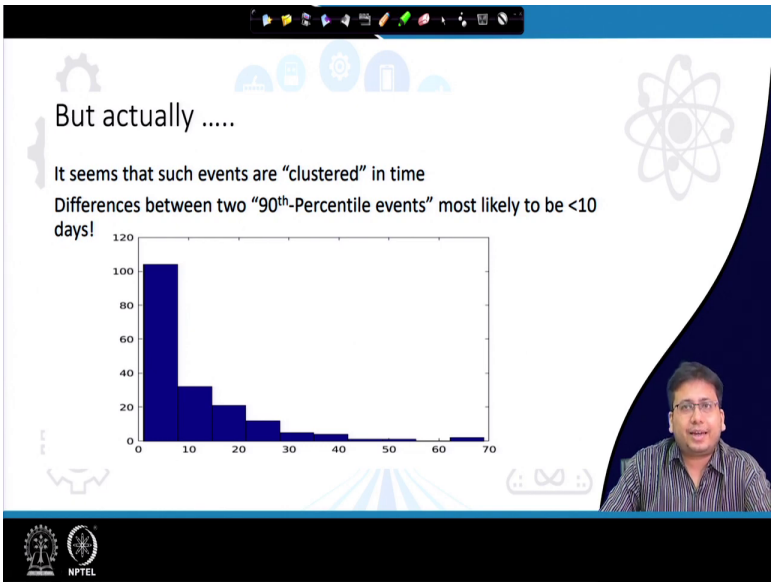
So, now once it happens today I may think that this kind this event happened today. So, much rainfall happened today in Kharagpur. The last time this happened was let us say 20 years back or 35 years back, these kinds of statements we see being made when there something very unusual happens and corollary to that is when do we expect it to happen next.

So, like the general idea is that the common events may happen soon afterwards. But the rare events like they because, they are rare it is like it is unlikely that once they have happened it is we may think that it is unlikely that it will happen again. It is something like saying that lightning does not strike the same place twice. So, the return period is defined as $1/p$ where p is like the event probability.

So, like this is something like the it follows from this $1/p$ which something that follows from the geometric distribution, which is related to how many times I must toss a coin before with bias p before it gives me a head ok. So now, if we consider a 90th percentile event the probability of a such an event happening is $1 - 0.9 = 0.1$.

Why because in 90% of the day the value observed will be less than that 90th percentile value right. So, like the probability of such an event happening which is at least as heavy as the 90th percentile like it is only 0.1 and the return period is $1/0.1$ which is 10 days. So that means that once this kind of an event happens a 90th percentile event, so, in case of that Kharagpur rainfall data set that 90th percentile was 25 millimeters. So, like it is like saying that if on a given day that there is 25 millimeters of rainfall we may expect such a thing to happen again after 10 days ok.

(Refer Slide Time: 20:04)

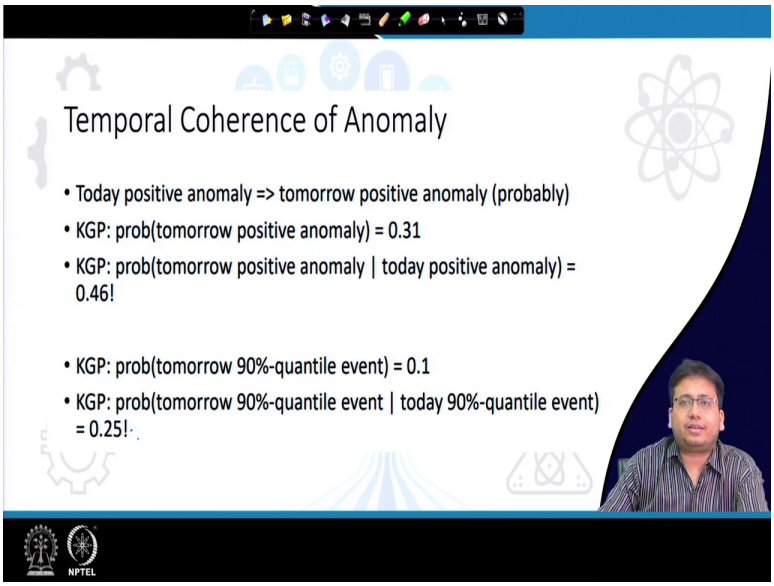


However, this kind of a thing actually does not happen. What happens is that extreme events are often clustered in time.

And the differences between the two 90th percentile events is most likely not going to be 10 days as was calculated here. But something like it can have be much smaller. So, here is actually a histogram of the difference between successive like 90th percentile events like that. So, you can see that once again the bulk of the distribution lies or the bulk of the probability mass actually is lies in this range say 1 to 8 or something like that.

So, most of that means, in vast majority of the cases the return the actual return period between two 90th percentile events is much less than the expected return period.

(Refer Slide Time: 21:03)



That means, what? That means, that these kinds of extreme events tend to be clustered in space. So, we can actually calculate this as follows. So, let us say let us just calculate that tomorrow there will be a positive rainfall anomaly.

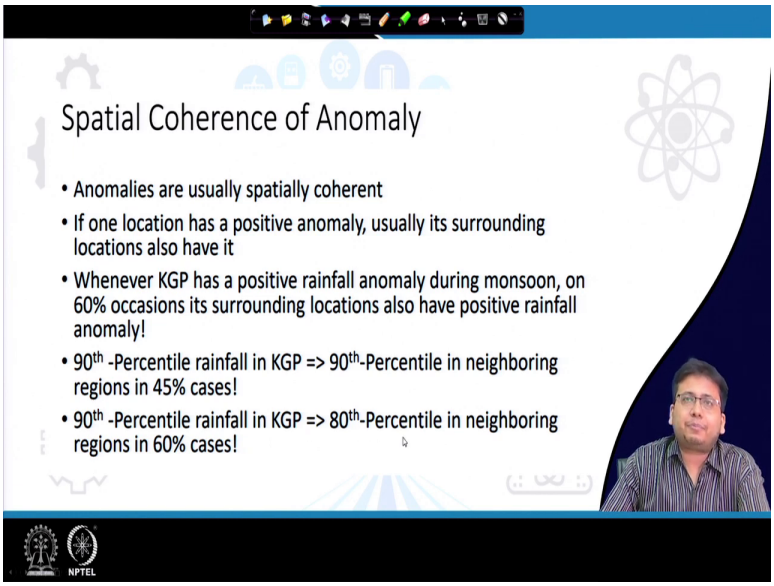
So that means, more than 9.13 millimeters of rainfall. The probability of that is of course 0.31. So, we have seen that in the data there are some 573 or positive anomaly days which is 573 out of 1830 is 0.31. So, there is 31% chance that tomorrow there will be a positive anomaly.

But if I already know that today there has been a positive anomaly, then the probability that tomorrow there will be a positive anomaly goes up to 0.46. So that means, positive anomaly is tend to be clustered in time, if today there is a positive anomaly most likely tomorrow I mean that it is likely that tomorrow also there will be a positive anomaly that is how these kinds of things work.

So, the concept of return period I mean it does it is not actually very practical, it does not actually tell us much I mean like it only gives us an expected value of when that event is likely to return. But typically it is not like that. Now if we like similarly instead of considering the positive anomaly if we consider the 90^{th} quantile or 90^{th} percentile event. So, the probability of that is of course 0.1 just by definition.

But if you know that today a 90th quantile event has happened then the probability that such an event will again happen tomorrow shoots up to 0.25; that means, like from 0.1 it has gone up to 0.25. So, again it shows like the same effect that these kinds of events are usually clustered in time, not only in time but they may be clustered in space also.

(Refer Slide Time: 23:05)



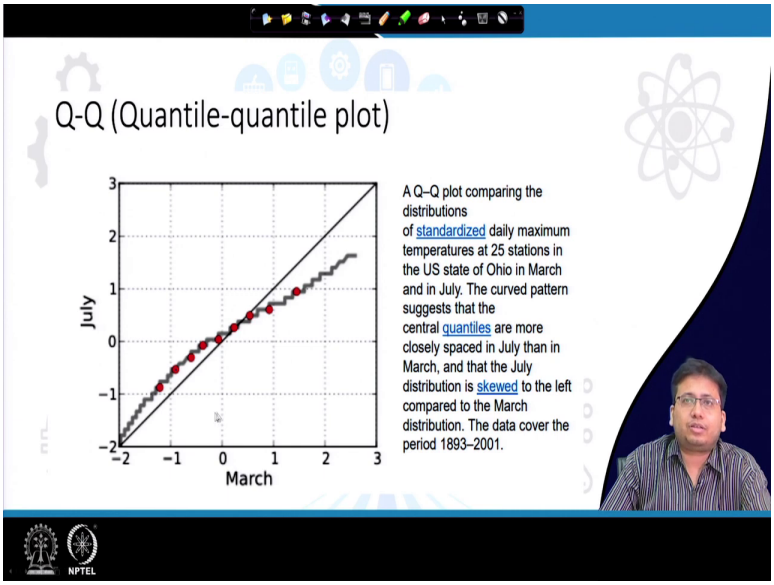
So, if one location has a positive anomaly usually it is surrounding locations also have a have such an anomaly. So, like an analysis of the data shows us if Kharagpur has a positive rainfall anomaly during on any day of monsoon, on 60% of the occasions it is surrounding locations also have a positive rainfall anomaly. So, the in the absence of any further information the probability of having a positive anomaly as we have seen is just about 30%.

For nearing locations it might be slightly more or less, but if we know that on a any given day Kharagpur is having a positive anomaly, then that probability shoots up from 30% to nearly 60% which is a huge rise. Similarly if you considered the 90th percentile rainfall events in Kharagpur which we saw is some 25 millimeters of rainfall or something like that.

It turns out that 45% of those days are accompanied by 90th percent rainfall in the neighboring locations also. Similarly and not only that in 60% of those cases the neighboring locations have

80th percentile cases ok. So, that this shows the spatial coherence of these kinds of extreme events and now this brings us to the concept of Quantile-quantile plots.

(Refer Slide Time: 24:34)



So, here basically we are comparing the like the different various quantiles of two different distributions. So, let us say for example, here like we are considering the rainfall distribution the in March at a given location and here we are considering the rainfall distributions in July at the same location. So, like the you may be surprised to see the values of -2, -1, etcetera.

I mean how can rainfall be negative, but so what they have done is, they have in fact, it is not rainfall sorry these are temperature. So, what they have done is they have standardized it which means that like basically they have removed the anomalies. So, here we can see that like this -2 is something like the lowest value and +3 may mean the may be the highest value.

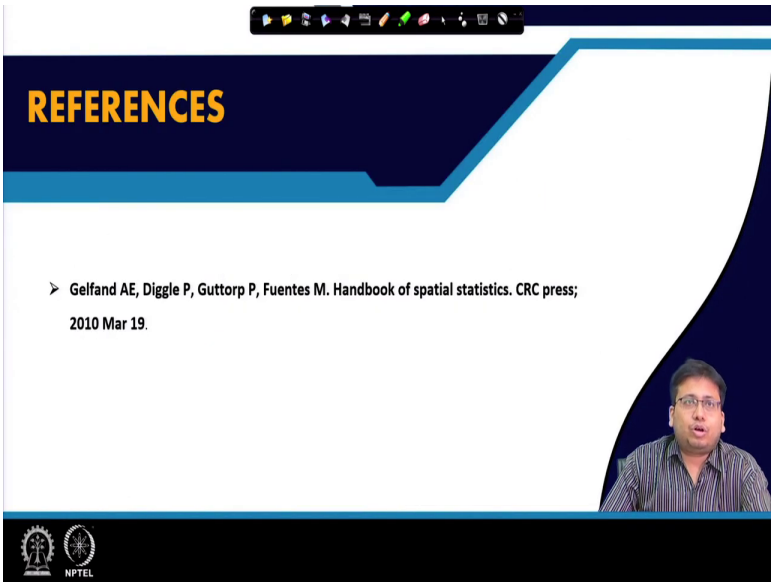
So, here what they are doing is that they are plotting the different quantiles in March against the same quantile in July. So, basically they have plotted something like the 50th percentile in March versus the 50th percentile in July or the 75 like along the $x - axis$ we may have the 75th percentile in March which may let us say that it is this value 2 and on the $y - axis$ we are plotting the corresponding 75th percentile in July which is something like 1.1 or something like that.

So, as you can see the if we consider the higher percentiles then march in March it is higher, but if you consider the lower percentage then in July it is higher. So that means, like if that like if you consider the let us say the hottest day in march which is something like the 99th percentile or something like that, that is very high that is like that may be something like 3.

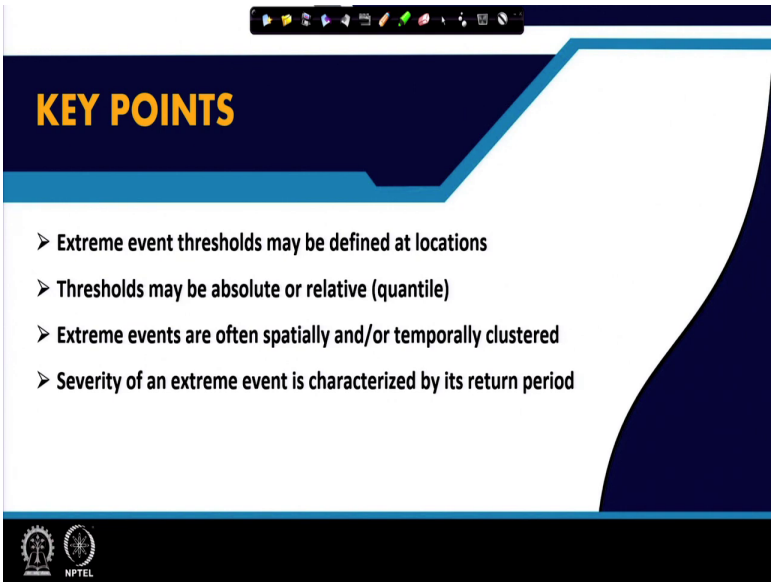
But in case like the corresponding value in case of July might be. In fact, it is not 3 it is something like let us say 2.7 right the end of this distribution some somewhere here. But the corresponding value the highest value in July is actually much less it is like something like 1.8, but it is quite the reverse if you consider the lower end of the spectrum.

If we so it is like saying that the hottest day in March is hotter than the hottest day of July, but the coldest or the colder days of march are cooler than the colder days of July. So, this kind of a Q-Q plot, it enables us to match the nature of these extreme events at different locations or at different time points as these cases.

(Refer Slide Time: 27:50)



(Refer Slide Time: 27:56)



So, like this sets up a very background for a an further discussion of these extreme value statistics. So, like actually the various probability distributions which we are familiar with they do not work well when we are dealing with exactly with the extreme values. And in the next lecture we will actually see how we can define specific probability distributions for these extreme events. So, the key points to be taken away is that the extreme event thresholds they may be defined at locations.

The thresholds may be absolute like something like $\mu + 2\sigma$ or relative which is the defined quantiles, the extreme events tend to be specially or temporally clustered and the severity of an extreme event is characterized by its return period. So, with that brings us to the end of this lecture, we will continue our understanding of these extreme events in the next lecture as well, so till then bye bye.