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 - 07
Extreme Value Theory

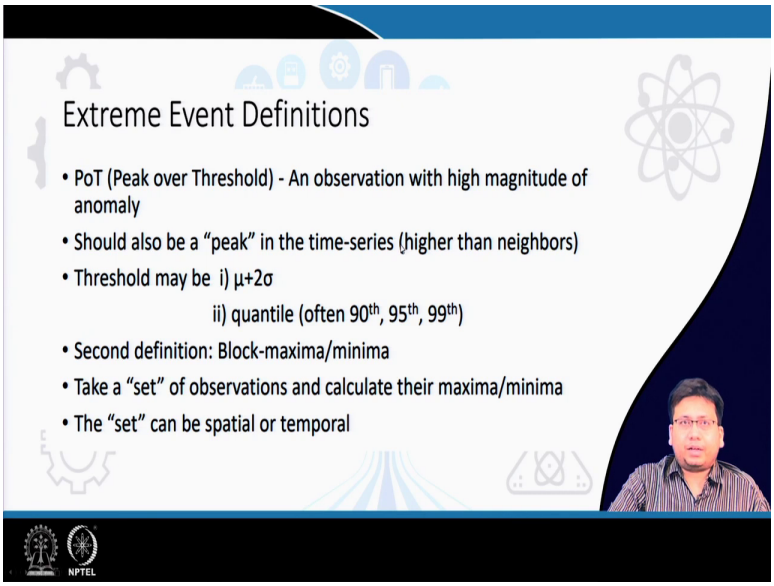
Hello everyone. Welcome to module 7, the module 1 of this course on Machine Learning for Earth System Science which is on Spatio-Temporal Statistics. Today is the lecture 7 of this module and today we are going to deal with Extreme Value Theory.

(Refer Slide Time: 00:43)

So, if you remember in the previous lecture, we had already been discussing about extreme values of like the geophysical variables like the precipitation and temperature and so on. We had defined quantiles and things like that and how to define extreme events in terms of such quantiles.

Now, today we will learn a few more concepts; probabilistic estimation of return levels and return periods. So, return levels and return periods are concepts which we had already studied in the previous lecture, today we will see how to estimate them from using probability distributions.

(Refer Slide Time: 01:27)



And, the second thing which we will learn today is generalized extreme value distribution. So, first of all let us just come back to extreme event definitions. So, in the previous lecture we had considered setting some kind of threshold on the value of the geophysical variable or rather on the anomaly. If the anomaly is above certain threshold, we will call it as a like as a positive extreme and if it is below certain negative threshold then we will can call it a negative extreme.

So, this kind of definition is sometimes known as the peak over threshold definition of extreme events. So, like where an observation with a high like where the magnitude of the anomaly is high. And, like if we are plotting the time series of the variable then the extreme events which we are talking about they can they also usually correspond to peaks or troughs of the time series like that I mean which are like either higher than their neighbors or higher than some threshold.

So, in the sense it is like a local and global maxima or minima. Now, the thresholds which we are talking about, those thresholds can be either like absolute like say something like $\mu + 2\sigma$ or it can be in terms of actually it can be some like above some fixed value or it can be like something

like $\mu + 2\sigma$ or it can be in terms of these quantiles. So, like for example, the 90^{th}, 95^{th} or 99^{th} quantile percentiles.

The there is a different definition of extreme events also that of the block maxima or minima. So, the this approach is that instead of considering individual observations, we consider sets of observations and calculate their maxima or minima depending on the kind of extremes we are looking for positive or negative. And, this set of observations it can either be a spatial set or a temporal set; that is either we can consider the observations over like a certain period of time or over a certain number of locations.

(Refer Slide Time: 03:44)

When we are dealing with block wise extremes in this case say for example, we consider the seasonal maximum precipitation over Kharagpur. So, in that case, instead of dealing with all the observations, we now deal with only the maximum values; that is to say in a any given year we consider we do not consider the values of daily rainfall, instead we for every year we keep just one value which is the maximum daily rainfall in that year.

Of course, it need not be the blocks which we are defining need not corresponding to 2 years, they could also correspond to months, weeks or anything like that. But, like in this example, we are considering years. So, the advantage that this approach has is that there is less variance.

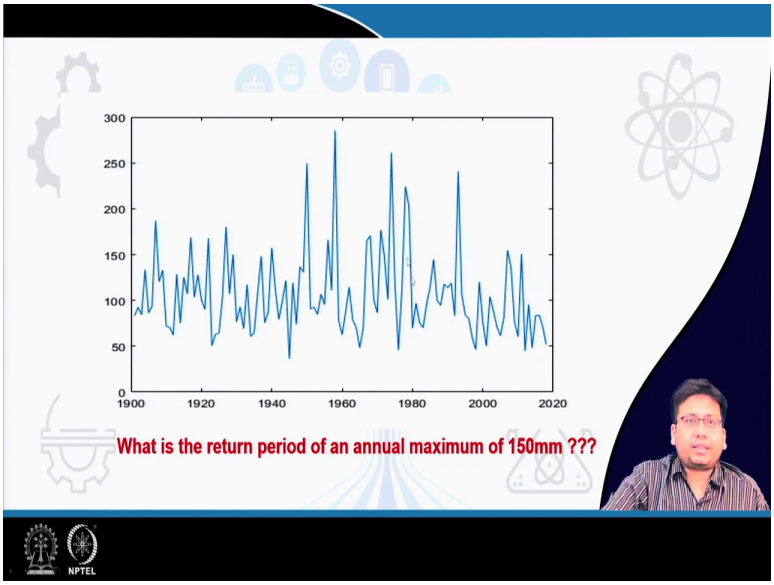
Like if we are considering data the variance will be much higher because, there will be many days in which precipitation will be absent totally that is completely dry days and there can also be days of say where there is some 150 millimeter of rainfall or so on.

In the previous lecture, if you remember the time series of 15 years that we considered, there was one day in 2007 where some 152 millimeters of rainfall had been recorded. So, that shows how much that the variance can be if we are dealing with daily data. But, if we deal with only block wise data the such variance is significantly reduced and in fact, larger is the block size, the smaller we can expect the variance to be.

However, the problem with this approach is that there is considerably less data that is instead of having a 8 15 years, we can have 1830 data points, if we are going for daily data. But, if we are going for annual maxima then for 15 years we will have only 15 data points. So, that is a significant reduction of these data points. As a result, any statistical quantity which we may want to estimate, they will be such estimates will be increasingly less robust ok.

Now, also the distribution of these values can will also significantly change if we consider the extreme values or these block maxima only instead of considering all the daily values.

(Refer Slide Time: 06:11)



So, this is actually the block maxima time series of rainfall over Kharagpur for 118 years from 1900 to 2018. So, its like in every year we are just recording the maximum rainfall that was recorded on any day of that year. Like say in this year for example, they later may be this is 1924 or something like that.

So, here we see that the day with the most with the maximum rainfall had only about 50 millimeters of rainfall, while in this year which might be say 1956 or 57 something like that, in this year the there was one day in which there was at least 290 millimeters of rainfall. So, that amount of deviation is there.

However, you like we can still say that the variance of this time series may be less than the variance of the daily time series which we saw in the other day. Now, based on this time series of annual maxima of daily precipitation; now I am asking you one question. What is the return period of an annual maxima of 150 millimeter? That is say on any given day sorry in any given year let us say there was a day in which the maxima the in which the rainfall was 150 millimeter.

So, what is the probability or what is the expected number of years after which this kind of thing will happen again? So, if you empirically look at this time series, you will see that there are a number of years where the maxima has exceeded 150 millimeter. Like for example, this year, this year, this year etcetera. In fact, if you like there are 20 years in this period in this 118 year period in which the annual maxima has been 150 millimeters or more.

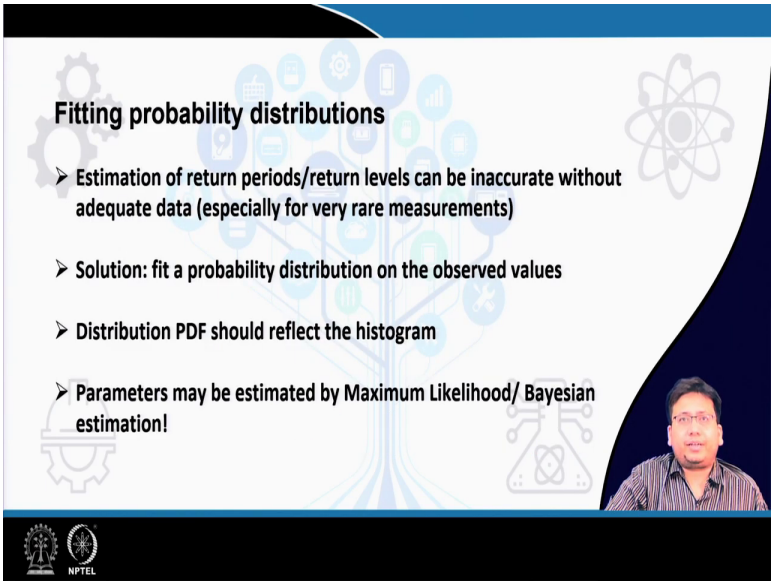
So, when I am talking about annual maxima of 150 millimeter in the context of extreme values or these return values, we do not mean specifically 150 millimeters; it we always mean 150 millimeter or more; even if even though we do not write the or more part you can like consider it as implicit. Now, how to estimate these kinds of return periods? So, in the last lecture what we were doing is we are trying to do empirically.

We were simply like for example, we were like for so, some particular values we calculated the number of days in which the rainfall had exceeded that value in that 1830 year period, I mean 1830 day period. And, accordingly we were setting its various percentiles like 19, this is 99^{th} percentile, that is 90^{th} percentile and so on. But, that kind of analysis we can do only if we have a lot of data that is like 1830, we can say is a reasonably large number of data points.

But, in this case as you can see the number of data points is significantly less, only 118 data points are there. So, the more the now the error here scales with the square root of the number of data points. So, like if as we want to reduce the error, we need to increase the number of data points quite significantly. So, what happens is that if we like in this kind of a situation when we are dealing with annual this block maxima.

Hence, we have significantly less data do going for by the empirical method of just counting how many times this event has happened in the in this period and just taking the inverse frequency ratio of that is usually not a great idea. Because, as I said the data is not enough to make a statistically significant or statistically robust estimate like that. So, what do we do? The answer is that we try to fit probability distributions.

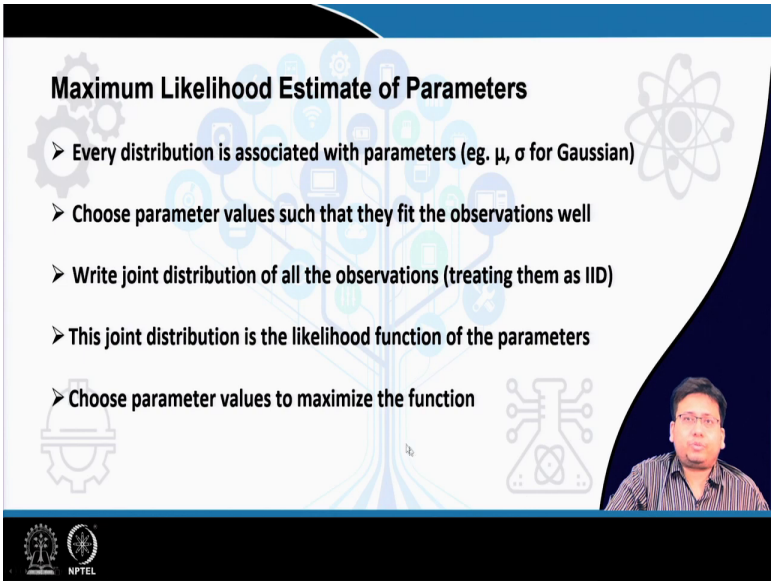
(Refer Slide Time: 10:21)



We like fit some kind of a distribution on the quantity of interest and then find the probability of the specific event that we are interested in like which may be that the daily rainfall exceeding some peak or the annual block maxima like exceeding some peak or something like that ok. So, the basic idea is that we fit a probability distribution on the observed values. Now, the distribution PDF that should reflect the histogram of the observations.

And, then the parameters may be of the distribution, they may be estimated using Bayes maximum likelihood or Bayesian estimation or something like that. So, whichever probability distribution we are fitting be the Gaussian distribution or gamma distribution or Bernoulli or whatever, it will have its own set of parameters which need to be estimated and that yes parameter estimation will be done from the data itself.

(Refer Slide Time: 11:20)



So, there are certain standard approaches for estimation of these parameters. The most well known is the maximum likelihood estimate. So, while I leave it for to you as a homework exercise to read more about how exactly maximum likelihood estimate is done, in case you do not know it already; I will just give you a the basic outline. So, the idea is that as we know every distribution is associated with certain parameters.

So, then what we need is we need to choose the parameter values such that the they fit the observations well. Note that we do not have the probability distribution to start with, what we have is only the observations. Now, we have to imagine the observations to be realizations of some random variables and then fit an appropriate probability distribution to the to that random variable. Now, choosing a probability distribution means two things.

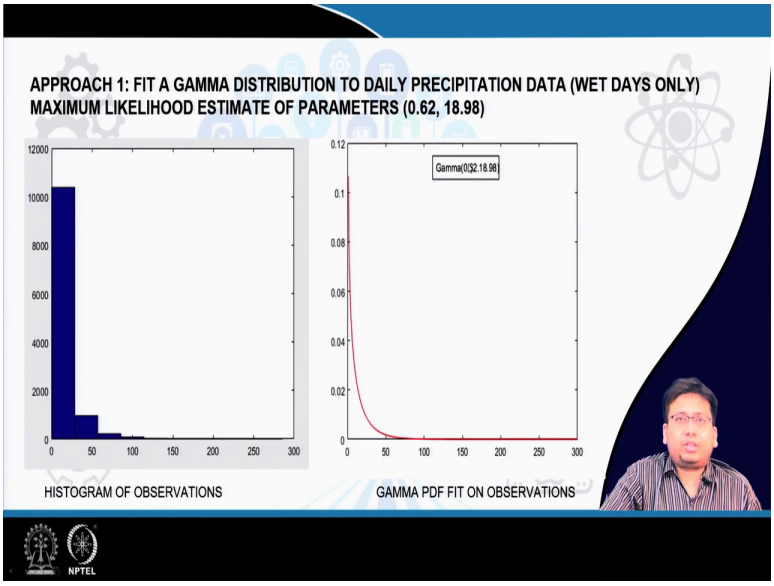
One is to choose the family of the distribution, is it Gaussian or gamma or whatever and the second is to estimate the values of the parameters. So, for the second purpose like what we do is we write the joint distribution of all the observations by treating them as IID and then this joint distribution is called the likelihood function of the parameters.

So, note that when we are writing the this joint distribution of all the observations, this like that such a joint distribution will include the values of the observations as well as the various parameters.

But while the values are known because they are observed, the parameters are so far not known, we are trying to estimate them. So, they are the variables of the of this joint distribution which we are writing. So, we call that joint distribution as the likelihood function of the parameters and now we will choose the parameters in such a way so, as to maximize this likelihood function.

So, we can write like once we have the function in terms of those parameters, we can just use something like we can write the partial derivative with respect to each of the parameters equate those to 0. We will get some simultaneous equations which we solve and get the estimates of the parameters.

(Refer Slide Time: 13:42)



So, this is called the maximum likelihood estimate. Now, in this case when we are the aim is to estimate the return value of the I mean the return period of 150 millimeters of rainfall.

So, what we first of all what we do is we plot the observations. So, like we so, for the time being forget about the block maxima part, let us say we are just dealing with the daily rainfall values. So, we now let us so, if we plot their histogram, this is what that histogram looks like. So, you can see that the low values have very high probability and the high values have increasingly less lesser and lesser probability.

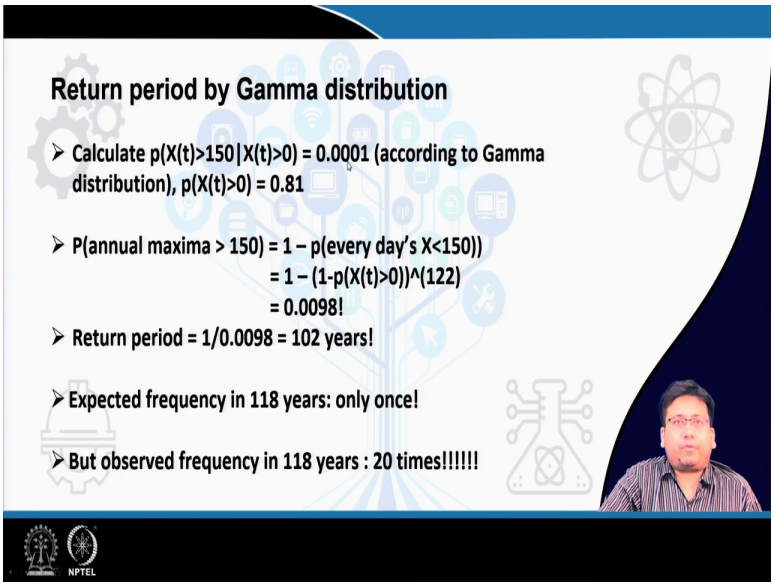
Now, this so, this is the PDF. Now, which what if we want to imagine that these observations are actually realizations of some random variables, we should choose such a distribution who whose PDF is approximately has this kind of shape. So, what is to mind is the gamma distribution. So, the like if like we know that the gamma distribution has this kind of step that like as you can see, this is a typical PDF of a gamma distribution.

You can see the similarity between these two curves and it is like usually gamma in the literature, gamma distribution is considered as a good distributed as a suitable distribution for this daily rainfall. The only catch is that a gamma distribution is not defined for like for 0 values, that is the support of the gamma distribution is like 0 to positive infinity. But 0, but specifically the point 0 is like it the PDF is undefined at the point 0.

So, what we will do is, we will like consider only the wet days that is where the rainfall is non-zero, we will estimate the parameter like that is we will fit a this kind of a distribution, the gamma distribution to only the wet days where there is non-zero amount of rainfall. So, like you can say that the amount of rainfall on any given day will be like modelled as follows on like, first we will choose whether it is a dry day or a wet day.

If it is a dry day then the rainfall amount is automatically 0 and if it is a wet day then the amount of rainfall we say it follows the gamma distribution ok. So, like by focusing the gamma distribution only on the wet days, we calculate the maximum likelihood estimates of the gamma parameters as these two 0.62 and 18.98. So, like the gamma distribution it has two parameters, the shape and scale parameters. So, those like these are the values which we estimate from the data by maximum likelihood.

(Refer Slide Time: 16:47)



Now so, using this gamma distribution let us do the mathematics and try to estimate that return the return period of 150 millimeters. So, first of all like when we are talking about the that the annual maxima or in any in a particular year is 150 millimeters what; that means, is that there is at least one day in that year where there has been more than 150 millimeters of rainfall right.

$$P(annual\ maxima\ > 150) = 1 - p(everyday's\ X < 150) = 1 - (1 - p(X(t) > 150))^{122}$$

So, the probability of such an event is of course, the complement of the event that every single day has a rainfall of less than 150 millimeter. Now, it is the second thing that we will try to calculate using the distribution above. So, the event that like every days rainfall is less than 150 millimeters, this can be decomposed into that is 1 days rainfall less than 150 millimeters and raised to the power of the number of days that is every day we treat as independently.

So, that the this joint distribution the it just becomes the probability of the individual distributions that is the probabilities for each and every day. So, remember that there are 122 days only in the season. So, what we like so, what we need to write is the probability of a rainfall in a particular day is less like is less than 150 millimeter right. So, sorry for the typo here, this will be 150 not 0.

So, now this the probability that the rainfall on any given day is like above 150 millimeter, that can be calculated using this kind of a gamma distribution as follows the probability. So, first of all the gamma distribution is remember that the gamma distribution is talking only about the wet days, where there is non-zero rainfall.

So, the gamma distribution gives us this probability, that the probability on any given day the rainfall is above 150 millimeter, given that it is a wet day ok and that probability comes out to be 0.0001, that is if you simply plug in these parameters to the gamma PDF, this is the value which you will get.

So, I leave it as a homework exercise for you to do this calculation once again. So, you can do it in MATLAB or Python. There are well known libraries for that, you do not have to write any function, you just have to that you just there is just one function which you need to call and provide it with these parameter values and it will give you the this probability.

Now, the probability of a dry day that turns out is sorry a probability of a wet day that turns out to be 0.81; that means, of the 18th of the days that we considered during this period whatever is the number of days that is 81% of them are wet; that is the rainfall is non-zero, only 19% are completely dry days.

So, the probability that on any day the rainfall is above 150 millimeters is this times this right and then. So, what we and the probability that any days rainfall is less than 150 millimeter will of course, be 1 minus that quantity right. And then so, we calculate the 1 minus that quantity and then raise it to the power of 122; 122 coming because there are 122 days in every year from June to September.

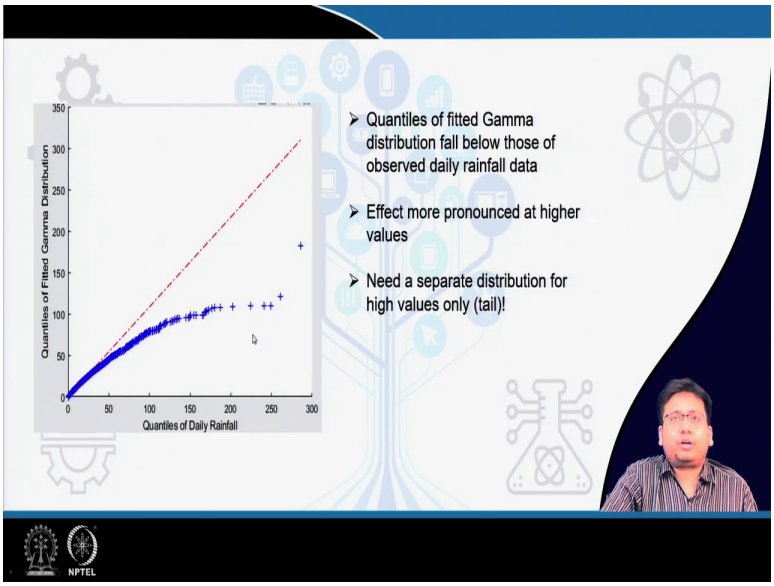
And, after cal doing this calculation the value the probability value we get is this thing 0.0098, again you should cross check these calculations. So, the; that means, what; that means, the $P(annual\ maxima\ > 150) = 0.0098$ and according to this model which is based on a gamma distribution on the daily rainfall amounts.

Now, based on this if you want to calculate the return period, this will be like as we saw in the last days last lecture that is the just the reciprocal of this probability which and it turns out to be

102 years; that means, this kind of an event will recur after 102 years each; that is nearly 100 years. But, here in this case we are considering a 118 years of data right.

Now, so within that 118 years if the return period is 102 years then of course, the in that period the expected frequency of this kind of an event that is annual maxima exceeding 150 millimeter, this will happen just once right. But, if you consider the actual data which we have that thing is such an event actually is found to happen at no less than 20 times.

(Refer Slide Time: 22:23)



So, the what the models prediction is horribly of the mark, it has said that this kind of event will happen only once, but it is found to be happening 20 times which. So, and if we the problem is lies in the quantiles of the distribution which we just fitted namely the gamma distribution. So, if you consider the QQ plot between. So, these are the actual observed observations of the daily rainfall over that period.

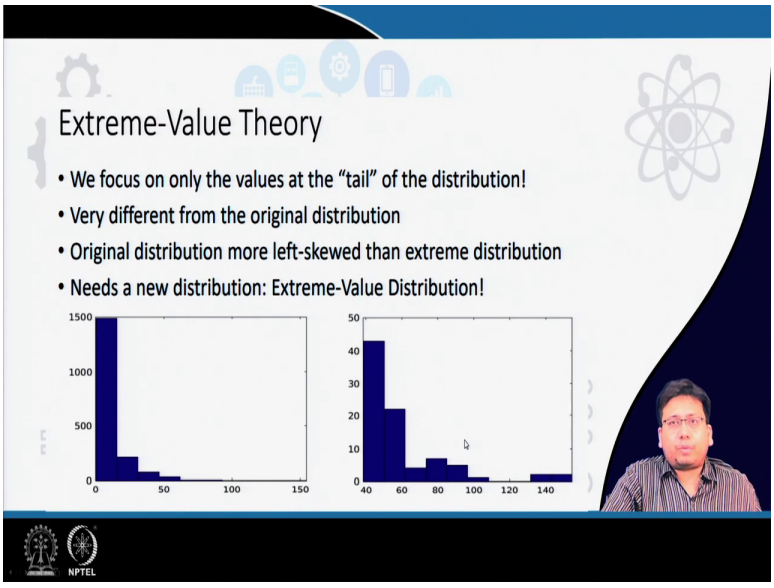
And, these are the samples drawn from the distribution which we just constructed the gamma distribution, we fitted the gamma distribution with those estimated parameter values. So, if you see the QQ, the idea like if the two distributions are perfectly in sync with each other or rather if the distribute, if the fitted distribution perfectly matches the data, then we expect the quantiles to follow this line.

But, if we find that it the plot follows this line only for like only for a certain period after that we find or rather only for low values both are matching. After that we find that like the fitted values, they are they just fall off the mark that is let us say this 200 millimeters. So, like this 200 millimeters may have a certain quantile a according to the fitted distribution. Let us say it is the 98th percentile or something like that or 95th percentile maybe.

But, the as the corresponding percentile in the fitted distribution with the 95th percentile is only about 100 millimeters. So, we and for the even higher values like say 250 millimeters and so on, we see its like the same case only. So that means, the probability distribution which we just fitted is unlikely to predict the kind of high values or extreme values which are actually observed ok.

So, the that means, what? That means, the probability distribution that we fitted the gamma distribution it fits the lower values quite well, but the higher values it is not able to fit. So, like the what is the need of the hour is to have a separate distribution only for the high values which are known as the tail of the distribution.

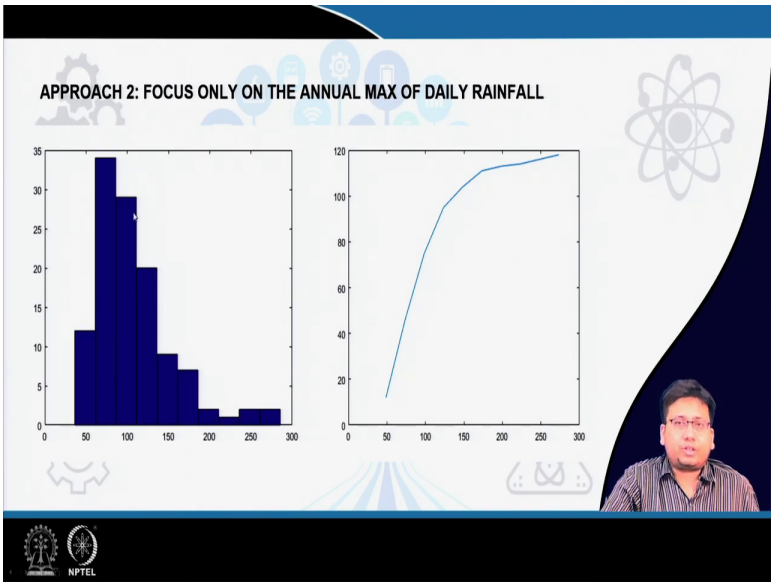
(Refer Slide Time: 24:39)



And, this is where the extreme value theory comes in. Now this so, in extreme value theory we focus only on the tail of the distribution and this distribution is usually very different from the original distribution. The original distribution is often very much skewed on one side, like if you

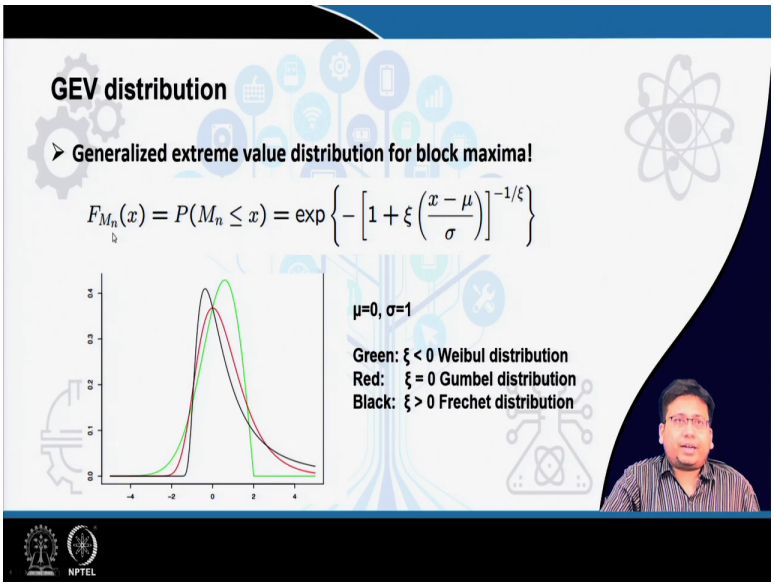
see like the original distribution the of daily rainfall it looks somewhat like this. But, if you consider something like a block maxima, the that is what like we like we may see a very different distribution.

(Refer Slide Time: 25:09)



So, if you like in this specific case for those 118 years, if you consider the block maxima of annual rainfall and like if you calculate the the histogram of that, this is what that histogram looks like which as you can understand is very different from the from the histogram of daily rainfall which we got here. You can like so, you can see this is very much left skewed. This on the contrary is not quite like that and this is the here what I am showing here is the CDF of that of the corresponding quantity.

(Refer Slide Time: 25:49)



So, now what we are going to do is we are going to suggest a new distribution which specific which is specifically meant for this block maxima. This is known as the generalized extreme value distribution. So, let us say that this x is the block or let us say that this quantity M_n, this like M is stands for some particular variable of interest and n stands for the block size.

So, what we are being saying is that the $P(M_n \leq x)$ follows this kind of a distribution. So, the this generalized extreme value distribution is derived this expression of this PDF is derived in a certain way, that is starting from the individuals or assuming some kind of an exponential distribution on the individual observations, we come to an like asymptotically we come to a this kind of a distribution over the maxima of the of a block of distributions and this is what that distribution looks like.

So, here you can see that there are basically three parameters; the μ, σ and this ξ. So, the so, like once again these are like they are the location shape and scale parameters, the μ is called the location parameter and this σ is the scale parameter. So, now what we can do like what we can do is when we have a set of observations, we can actually normalize them in a certain way.

So, that this μ becomes the becomes equal to 0 and this σ becomes equal to 1. This is a not identical, but analogous to the case where the where we have a normal distribution and we like

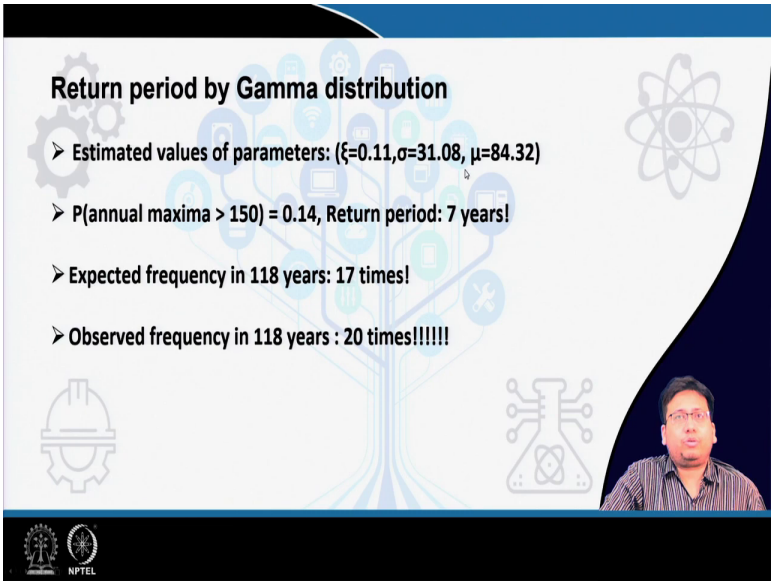
just generalize it and bring it to the standard normal distribution which has mean 0 and variance 1.

Now, the more interesting thing is this quantity ξ, this is in a sense it is like the shape of the distribution. So, if this ξ, if this $\xi<0$, then that specific case is called the Weibul distribution, whose PDF you can see here. So, as you can see this is actually a bounded maxima that is beyond a particular value there is a particular there is a fixed value above which there is 0 probability of occurrence.

Then, there is if $\xi=0$, then it is this is what is known as the Gumbel distribution. So, it is this that red curve. So, this is a situation where it asymptotically goes to the where the PDF asymptotically goes to 0, that is like as we go to higher and higher values this the PDF become like becomes smaller and smaller and ultimately converges to 0 in the limit. And, the last case the black curve where $\xi>0$, that is called the Frechet distribution.

So, here like this is called the heavy tail distribution where like even the very high values of this block maxima they continue to have a reasonable probability ok. So, now, like as I said these are the these three things; the location, shape and scale parameters are the parameters of this generalized extreme value distribution.

(Refer Slide Time: 29:22)

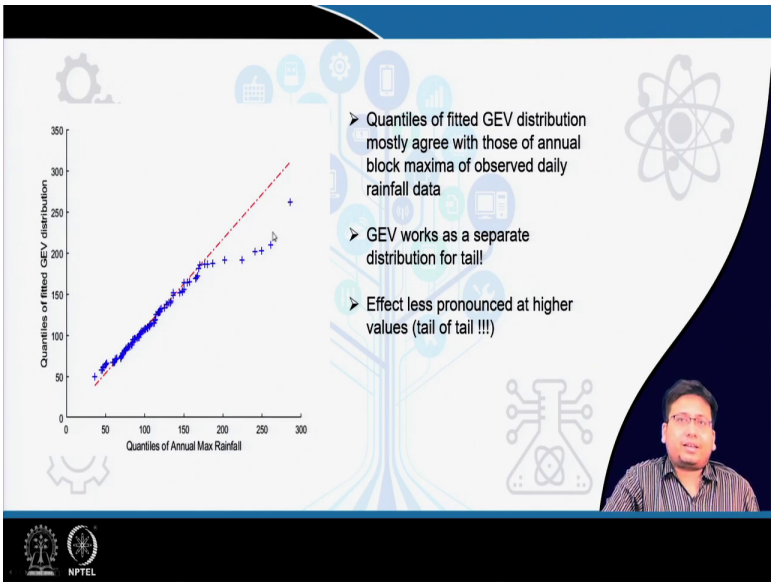


So, just like in the previous case, here also we can estimate those parameters using the maximum likelihood estimate. And, here this is the these are the maximum likelihood estimates which we have got in this particular data set, the 118 year data set. So, according to this distribution once we have fitted these parameters, we can calculate the probability that the annual maxima is going to exceed 150 millimeter.

And, that probability turns out to be 0.14 and its return period is then just the reciprocal of 0.14 which is about 7 years. That means; the expected frequency of this event that is annual maxima exceeding 150 millimeter, the expected frequency of such an event in 118 years is $118/7$which is 17 ok. So, we can expect that in the 118 year period such an event will happen around 17 times, like while the observed frequency is 20 times which is actually pretty much matching.

Now, compare it to the previous case of the gamma distribution on the daily values, where we had got which had predicted only once. So, from once we have come to 17 times of course, 17 times is still lesser than 20 times, but compared to what we have earlier, now we are doing much better.

(Refer Slide Time: 30:43)



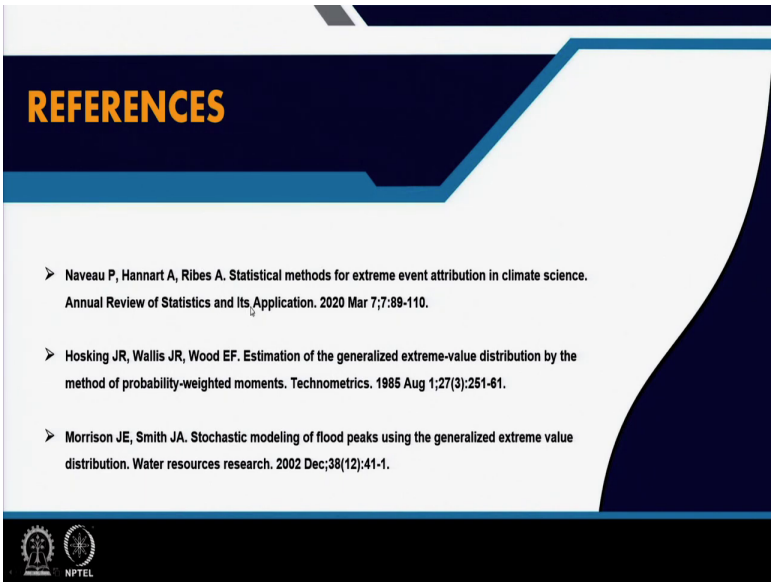
And, if you consider the QQ plot now like again just comparing the quantiles of the observed dist like observed values and the quantiles of values sampled from the this fitted distribution. Again,

we see that most of the part of the is like there is perfect agreement only like some very large values like we see that the it is underestimating the probability, the that is basically the effect we saw here; that is why instead of while the actual frequency was 20 times, while it predicted 17 times.

Like if we if instead of considering 150 millimeters, if we had considered an event of a threshold of say 100 millimeters or 120 millimeters; then probably we would have come closer to what is predicted by the model. But, if we had considered an even higher threshold like say 180 or 200 millimeters then probably the there would have been a bigger mismatch of the frequencies.

So, here you can see that even the GE, the GEV model is like slipping off, the I mean the quantiles of the GEV model distribution are slipping off near the what we can say the tail of the tail. So, like there is always the possibility of like focusing on the tail of the tail and trying to fit another distribution on it, but that is a different matter.

(Refer Slide Time: 32:10)



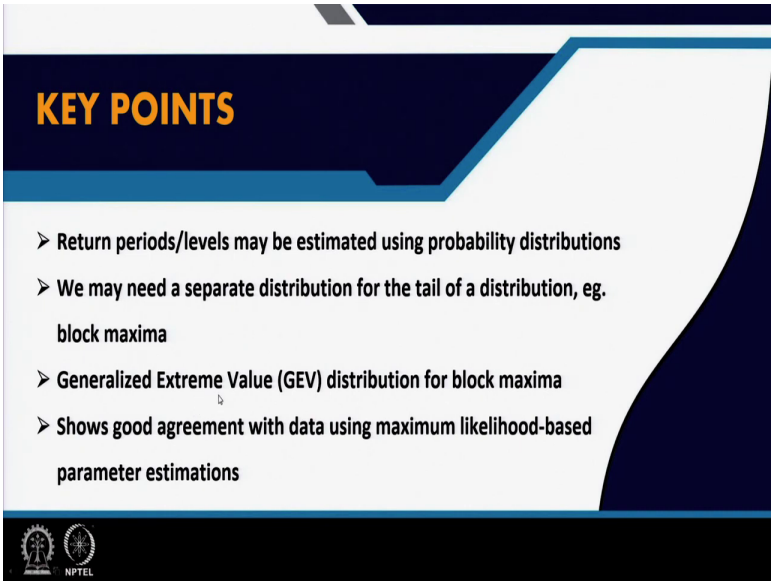
So, this concept of extreme values is quite important in the domain of earth system science. And, it has been used several times or in several research papers related to say like rainfall and climate, temperature and so on. Especially, in quantifying the impacts of climate change where

the I mean, one of the expected impacts of climate change is that the extreme values will become more and more frequent, that is their return periods will decrease.

An event that earlier could have been expected to happen once in 100 years, that is which had a return level return period of 100 years, now may its return period may reduce to only 10 years or something like that; that means, it can become about 10 times more frequent and things like that. So, there are many papers on this matter, I especially recommend you to go through this the first tutorial. So, this is some kind of a review paper or you can consider it as a tutorial.

So, this Professor Alexis Hannart, he is an expert on this extreme value theory and there are also other experts like Professor Richard Smith of University of North Carolina and others. So, I like I recommend you to go through their work.

(Refer Slide Time: 33:34)



So, the key points of this lecture are that the return values or the return periods they may be estimated using probability distributions. However, we may need a separate distribution for the tail of the distributions. And, one such distribution is the generalized extreme value distribution. This is suitable for block maxima and it shows good agreement with the data for maximum likelihood based parameter estimations.

So, apart from the so, we consider the GEV distribution for block maxima, remember that we also consider the peak over threshold definition. So, in that case also there are other specialized distributions called the generalized pareto distribution. So, you may I recommend you to go through them also.

So, with that we come to the end of this topic of extreme values. We will continue on other topics of this model. Till then, bye bye.