Stochastic Hydrology

Prof. P. P. Mujumdar
Department of Civil Engineering
Indian Institute of Science, Bangalore

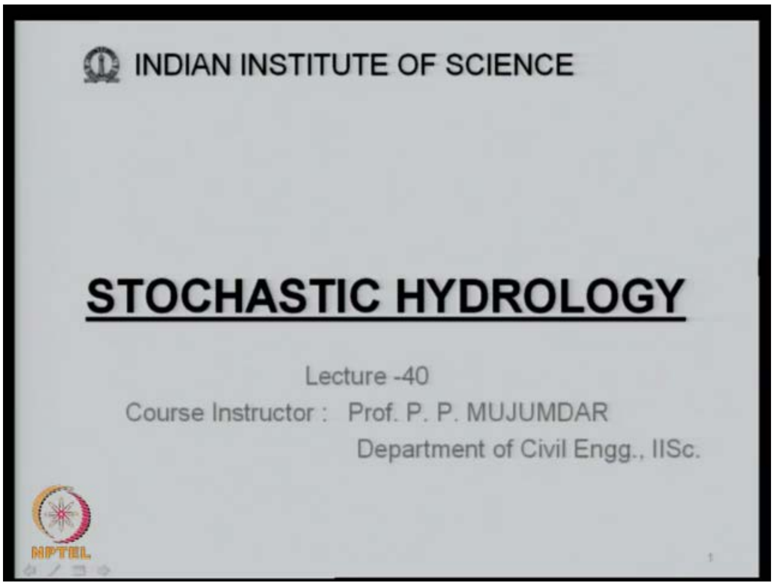
Module No. # 09

Lecture No. # 40

Summary of the Course

Good morning and welcome to this the 40th lecture of the course Stochastic Hydrology. This is the last lecture of the course, so what I propose to do in this lecture is to just run through the different portions that we have covered, different topics that we have covered in the course.

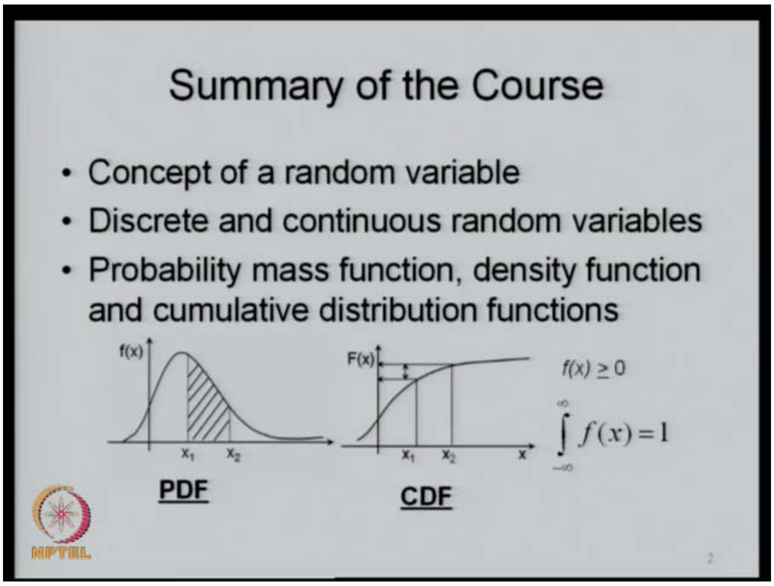
(Refer Slide Time: 00:38)



So, essentially this is a lecture covering the Summary of the Course. And then towards the end perhaps, we will spend just a couple of minutes on what other topics are of interest in the broad subject of stochastic hydrology; what is the scope and for further work what is the research that is going on in these topics and so on.

So, right from the first lecture until the last lecture, we will just go through the list of topics that we have covered, essentially just to recapitulate the important features, important topics that we have covered in this course. So, we started with the concept of a random variable, if you recall we defined the random random variable as a real valued function defined on the sample space.

(Refer Slide Time: 01:30)



And in fact, intuitively we understood the random variable to be a variable, whose value cannot be predicted with certainty, until it actually takes on a value. So, in that sense most of the variables that we deal with in hydrology or all random variables. For example, rainfall is a random variable, stream flow is a random variable, evapotraspiration is a random variable, humidity is a random variable and so on, so most of the variables that we deal with in hydrology or random variables.

Then, we define discrete and continuous random variables, those variables which take on only discrete values for example, number of rainy days is a discrete random variable. Because, it can take on values between values such as 0, 1, 2, 3 etcetera, only discrete values. And the random variables which can take on any value on a real line are called as the continuous random variables.

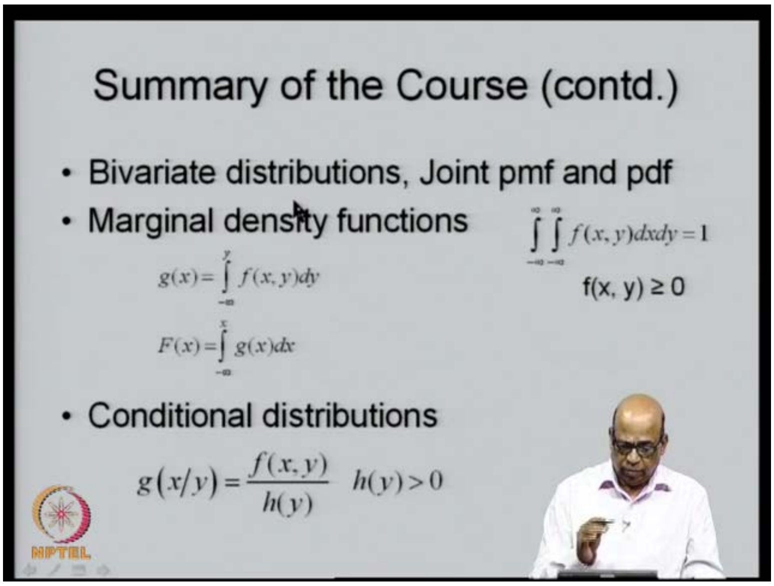
So, in that sense many variables that we deal with in hydrology for example, stream flow the the amount of rainfall during a particular period and so on, these are all real, that is continuous random variable. Then, we defined the probability mass function for a discrete random variable, and probability density function for a continuous random variable, and then we also defined the cumulative distribution functions.

For example, the PDF of a continuous random variable is shown like this, if you this is a probability density and this is a value that the on the x axis as the values that the random variable can take on (Refer Slide Time: 03:00). We denote, if you recall by capital letters the random variable itself, and by small letters the value that the random variable has taken for example, capital X can be stream flow and small x can be the value that the stream flow takes, for example 30 units, 40 units and so on.

So, from the probability density function, you obtain the cumulative distribution function by integrating the probability density function up to that point. For example, if you want the cumulative distribution function at x 1, you integrate from minus infinity to x 1, the probability density function to obtain this value.

So, capital F of x is minus infinity to x integration of f of x d x, the probability density function has the property that f of x is non negative and the integration between minus infinity and plus infinity is equal to 1, this is f of x d x.

(Refer Slide Time: 04:17)



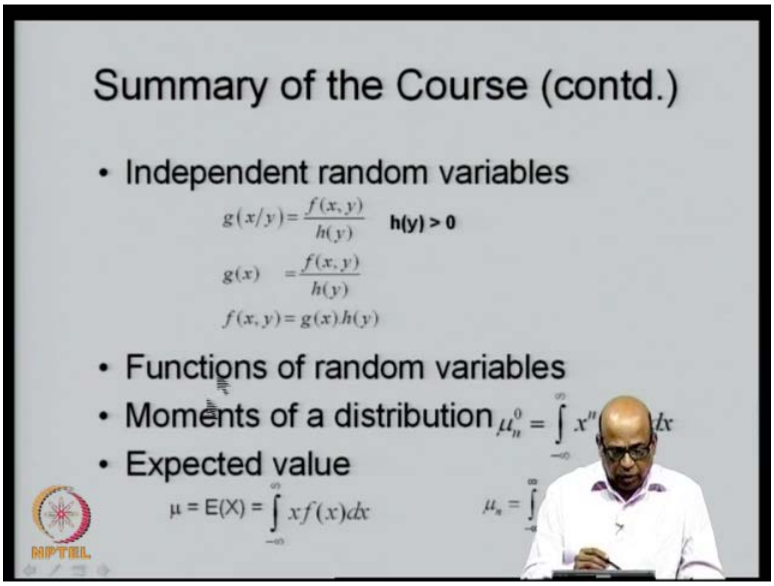
Then, we defined much the same way as we did for the univariate distributions; we defined bivariate distributions, where we are interested in joint behavior of two random variables. Let us say, stream flow at a particular location and rainfall in the catchment area or stream flow at a particular in a particular catchment with stream flow in a neighboring catchment and so on. So, we are interested in simultaneous behavior of two random variable then we call it as bivariate distributions.

The distributions that, we are defining for simultaneous behavior of two random variables. And the f of x y that is a joint PDF now, joint Probability Density Function for a continuous bivariate random variable, will satisfy these two conditions. Then, from the joint PDF, we obtain marginal PDF as minus infinity to y, f of x y d y; similarly capital F of x will be minus infinity to x g of x d x, so from this you are getting the cumulative distribution function.

So, given a joint density function you will be able to get the distribution of any of these variables x and y. Then, we defined the conditional distribution where we are interested in, the distribution of one variable conditioned on the other variable. For example, we may be interested in probability of x given y is equal to y, so such distributions we define it (()) conditional distributions.

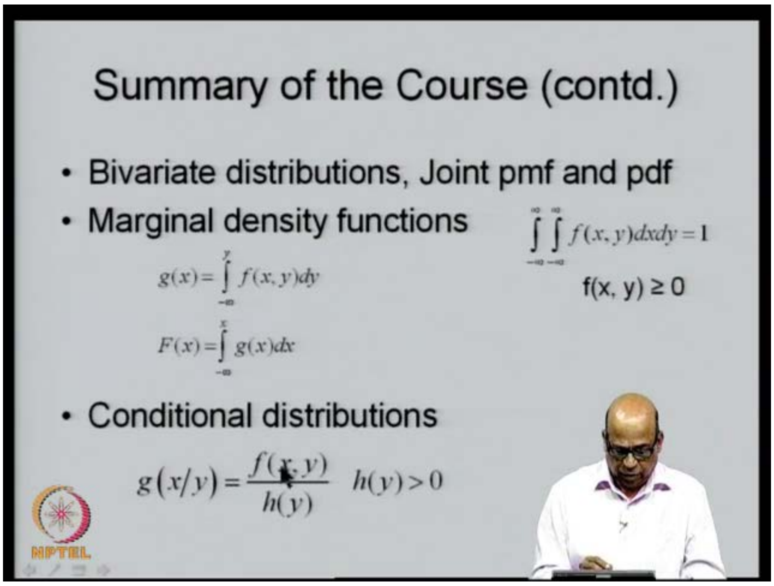
So, this is a conditional density, g of x given y, we define it as joint density f of x, y by the marginal density of y h of y, defined for h of y being positive. So, joint density, marginal density and then the conditional density, so these are all defined for bivariate distributions; similarly for multivariate distributions, when you have more than two random variables.

(Refer Slide Time: 06:38)



Then, we had an important concept are called as an independence.

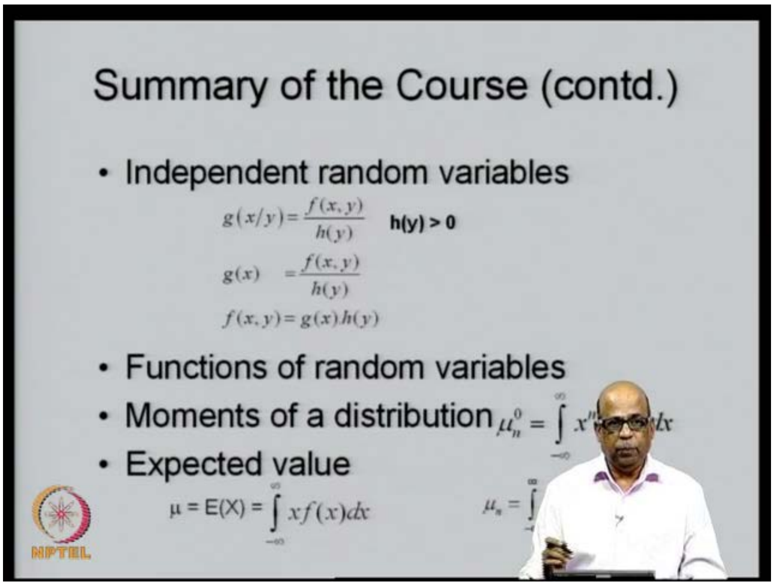
(Refer Slide Time: 06:44)



Now, if you have g of x given y, which is defined as f of x, y by h of y, let us say you have two random variables, which are independent of each other, what does that mean? It means that the value that one of the random variables takes is not influenced by the value that the other random variable has taken or (()) taking.

So, if you look at the conditional distribution g of x given y, we are talking about the density or the distribution of x, random variable x conditioned on the behavior of the random variable y. So, if x and y are independent, then irrespective of how the y random variable y behaves, the distribution of x should not change. That is when we call it as independent, call there are two random variables as independent random variables.

(Refer Slide Time: 07:40)



And therefore, when they are independent you have g of x, y by definition is f of x, y by h of y, defined for h of y being positive. If they are independent, then g of x, y can be x given y can be written as g of x, because that the distribution of x is independent of the distribution of y.

And therefore, g of x I write it as f of x, y by h of y, from this and therefore, f of x, y x, y is equal to g of x into h of y. So, two random variables are independent, if and only if their joint their marginal densities the I am sorry, I will repeat the two random variables x and y are independent, if and only if the product of the marginal densities is equal to the joint density itself.

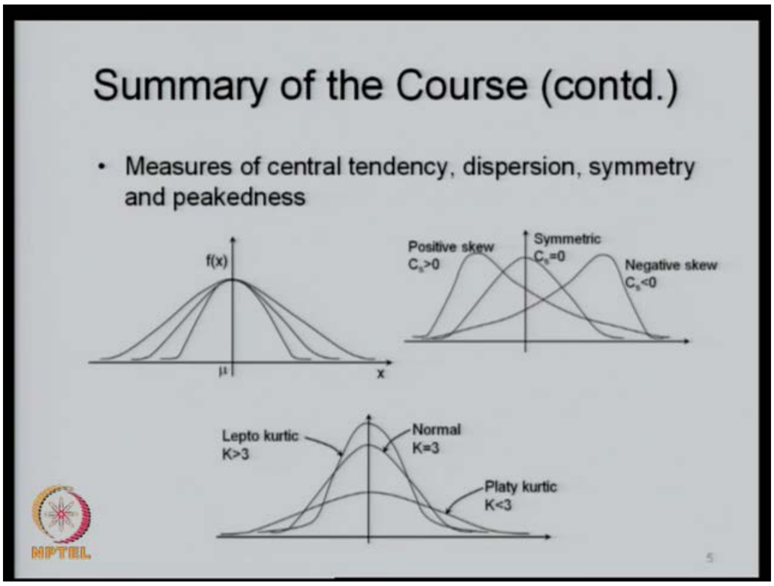
Now, this is an important result and we have seen some applications of that then we went on to define the distributions for derive, the distributions for functions of random variables, and typically we will be interested in sums of two random variables. If you recall the the example that we considered was, that you have a tributary coming and then you have a main stream, so that is a flow of capital X, the random variable is capital X there and the flow that is coming here is capital Y and downstream of the confluence, (()) interested in the distribution of the random variable.

So, the downstream of the confluence you have x coming from here, y coming from here and x plus y is downstream of the confluence. So, we will be we will be interested in often the behavior of the sums of two random variables, and that is what we did in the functions of random variables. We have derived, we have seen how to derive the distributions of sums of random variables not only sums, some functions of defined on random variables. Then we consider properties of random variables or on the distribution we defined moments for example, we defined a moment n th moment about the origin. So, mu n 0 here, we define it as x to the power n that is a integral x to the power n f of x d x, so this is a n th moment about the origin, and the first moment we about the origin, we called it as the mean or the expected value, so expected value is the first moment.

So, when you put n is equal to 1, you get minus infinity to plus infinity x f of x d x, that is an expected value of the random variable x; we often use the moments about the mean. So, mu n without this superscript here is the n th moment about the mean, so we define the n th moment about the mean as integral minus infinity to plus infinity x minus mu to the power n f of x d x.

It is, in fact, these moments that we consider to define several of the properties for example, the variance, variance will be the second moment and then the third moment, you have skewness and kurtosis and so on. So, all of those we defined based on the moments taken about the mean, moments of the distribution about the mean.

(Refer Slide Time: 11:40)



So, using these, we then defined the measures of central tendency dispersion, symmetry and peakedness. Now, all of these we define based on the moments for example, central tendency we looked at the mean, the expected value in fact, we also define the mode and the median. So, mean is defined as the first moment of the distribution about the origin, so the first moment is the expected value. Now, you can recall that, you may have several distributions all of which have the same mean, but their spreads may be different the dispersion the dispersion about the values, about the mean may be different from for different distributions.

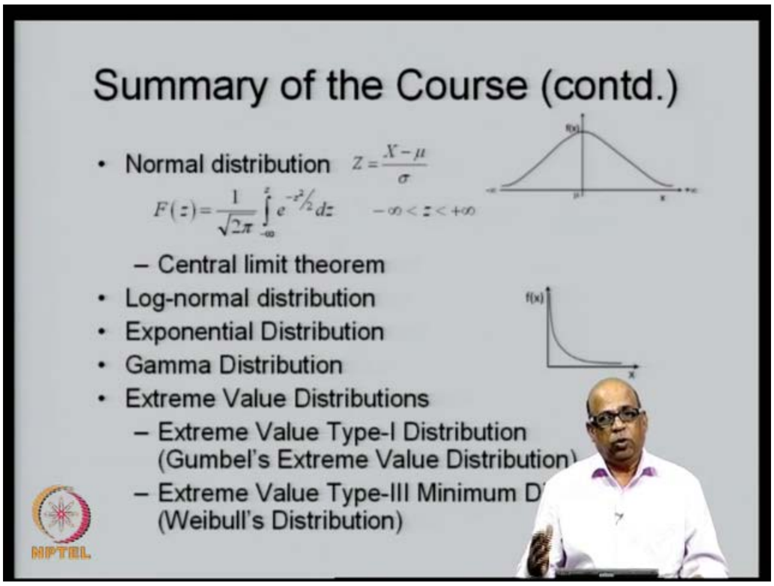
And therefore, we define measures of dispersion; typically we define the range which gives you maximum value minus minimum value the spread. And then we defined the variance which is in fact, the second moment about the mean and from the variance, we defined the standard deviation as the positive square root of the variance, and then a non-dimensional coefficient, called as a coefficient of variation.

Which is simply s by x bar or sigma by mu that is a standard deviation by the mean, both the standard deviation as well as the mean have the same units as the original variable. Then we looked at the skewness coefficient which is defined based on the third moment, third moment about the mean here, so based on the third moment, we defined the coefficient of symmetry or symmetric.

And we denote it as gamma or based on the samples it is denoted as C s, when it is estimated based on the samples it is denoted as C s; a perfectly symmetrical distribution will have a C s of 0. And then you may have positive skew with C s greater than 0, negative skew with C s less than 0, then we went on to the fourth moment and looked at the peakedness of the distribution.

So, you have a normal peakedness for the normal distribution and the kurtosis that we define, this k here is the kurtosis, k is equal to 3, and then if you have a flat distribution like this, we call it as a platy kurtic, where k is less than 3 and then you have lepto kurtic k greater than 3 with a very sharp peakedness. Now, all of these properties or the measures will be helpful in assessing what kind of distribution that you expect from the particular random variable.

(Refer Slide Time: 14:44)



Then we went on to study a few specific distributions, typically the continuous distributions and the most important distribution that we have studied and most frequently applied distribution, most useful distribution is a normal distribution or which is a symmetric distribution, it is also called a bell shaped distribution, it is called as a gaussian distribution, it is symmetric about x is equal to mu.

And then we defined the standard normal distribution, which is a very handy tool for calculating probabilities of normal distribution. And the standard normal distribution is a cumulative distribution of that is given by this expression, when we standardize this, the original random variable is X and this follows normal distribution. So, X minus mu over sigma, this is called as a standard normal variate or standard normal random variable.

And z it follows normal distribution with a mean of 0 and standard deviation of 1 and it has a cumulative distribution given by this, it is a standard normal distribution that we use to get the probabilities associated with the particular random variable, which follows normal distribution.

And tables are available for getting F of z for a given value of z, those tables we use to obtain probabilities on X, we have seen several examples of this, and the most important theorem associated with the normal distribution is a central limit theorem. Which essentially shows, says that if you have a large number of random variables, the sum of those random variables x 1, x 2, x 3 etcetera follows a normal distribution? The requirement, the original requirement there is that, these random variables are independent and identically distributed.

But, in the hydrologic applications we have seen that this requirement of identical distribution can be relaxed and then, when you have a large number of variables which are independent of each other. Then you can take the sum of those random variables to be following, approximately following a normal distribution; and this is an important result which we use (()) hydrology.

For example, (()) talking about monthly stream flows, monthly stream flows can be constitute as having being composed of daily stream flows, and daily stream flows are independent and you are adding them up to get the monthly stream flows, the monthly stream flows can be taken as constituting a normal distribution. Then we also studied log-normal exponential for example, the exponential distribution is like this for x greater than 0.

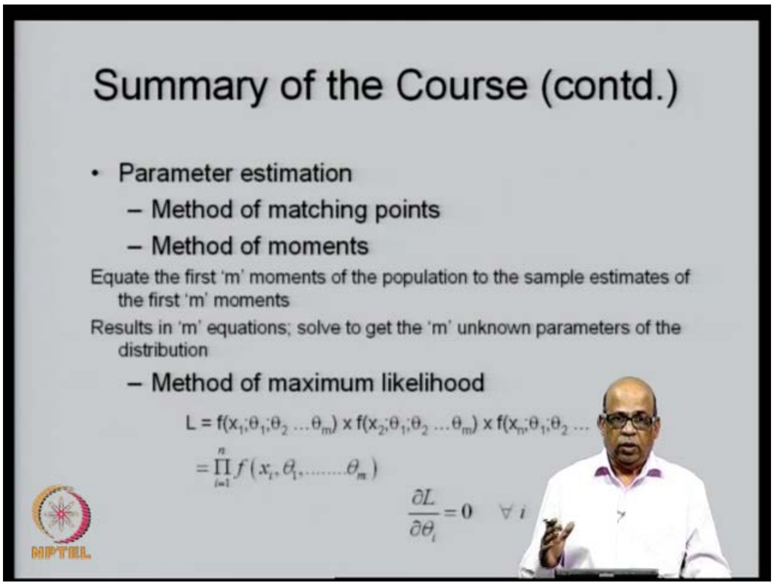
Exponential distribution has the memory less property, in the sense that the the moment generating functions, from the moment generating functions, we can show that exponential distribution is a memory less distribution. I think in this particular course we did not touch up on the moment generating functions, so for the time being, let us keep it aside.

Then we covered the gamma distribution which is in fact, a family of distributions it has shape parameters it has shape parameter and the size parameter; as you vary the size parameter and the shape parameter you get a different type of distribution all together, and therefore, gamma distribution consists of a family of distributions. And we have also seen how to use the tables of gamma distribution to obtain the probabilities.

In fact, we have seen that exponential distribution is a specific case of gamma distribution for a given a for a specified value of lambda and eta the two parameters. Then we went on to look at the extreme values, the extreme value type one distribution for example, we may be interested in how the maximum values of a series behaves for example, we are interested in flood peaks or maximum rainfall.

And then, we are also interested in the other end of the distribution that means, how the minimum values behave for example, low stream flows, in in the case of water quality, in the case of drought situations etcetera, we will be interested in this. So, we have seen the extreme value type I distribution, as well as extreme value type III distribution is called as a Gumbel distribution. It is often used in flood frequency analysis and extreme value type III is called as a Weibull's distribution, which is used at a drought analysis.

(Refer Slide Time: 19:39)

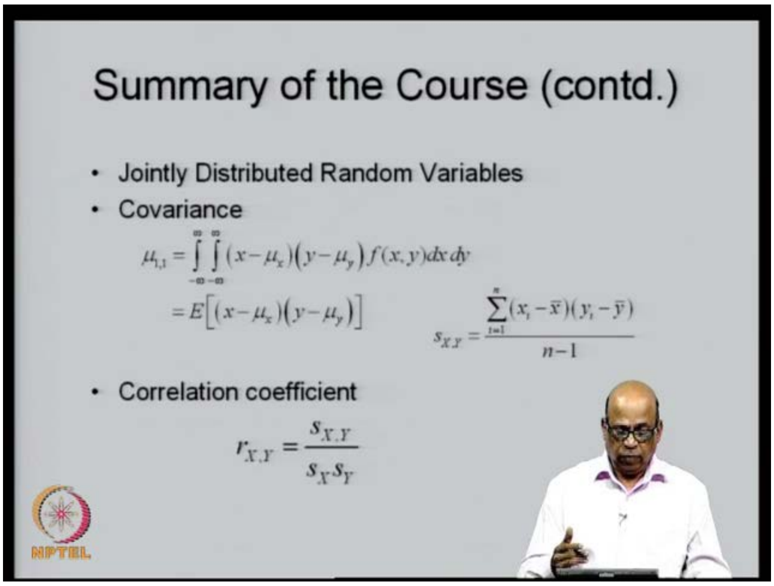


The parameter estimation for a given distribution, we looked at two methods in fact, three methods, method of matching points just look at the data, assess the probabilities then equate these probabilities to the theoretical probabilities that you obtain from the particular distribution and then obtain the moments, that is the method of matching points.

But, there are more rigorous methods which is method of moments, let us say you have two parameters, generate two moments from the distribution and then you get two equations, solve for these two equations you will get the two parameters. Then you have the method of maximum likelihood, where you define a likelihood function for example, f x 1, if you have m parameters theta 1 theta, 2, theta m etcetera, f x 2, x 1, x 2, x 3 etcetera, x n are the sample values.

So, you have a sample, based on the sample you defined a likelihood function like this where theta 1, theta 2 etcetera are the parameters that you want to estimate; then from the likelihood function, you obtain that set of theta 1, theta 2, theta 3 etcetera, which maximizes the likelihood function. So, this is how you obtain m equations in this particular case, solve the m equations to get the m parameters. So, this is how we obtain the moments using the method of maximum likelihood.

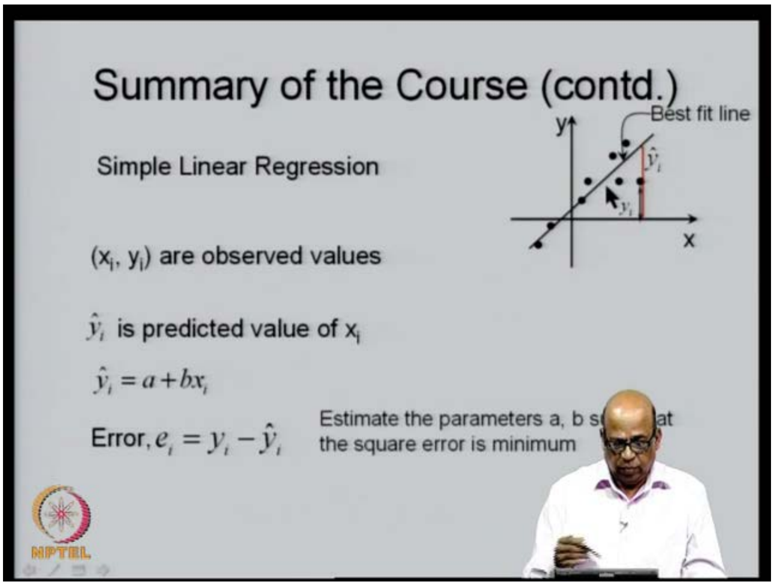
(Refer Slide Time: 21:16)



Then we consider the jointly distributed random variables we looked at the covariance for the univariate distribution, we had defined variance. So, for the joint distribution, jointly distributed random variables you consider the covariance which is a 1, 1 th moment. So, you defined mu 1, 1 as x minus mu x y minus mu y f x, y d x d y, and this we called as the expected value, this is in fact, the expected value of x minus mu x y minus mu y, it is called as the covariance. And from the sample we know how to estimate the covariance, this is how we estimate the covariance.

Then we defined again a a dimensionless quantity to show the dependence of one variable on the other, this is called as a correlation coefficient, again a very important concept that we use in stochastic hydrology r X, Y this is for the samples r X, Y is equal to S X Y by S X S Y. This is the sample estimate of the covariance, and this is a sample estimate for standard deviation of X, standard deviation of Y.

(Refer Slide Time: 22:30)

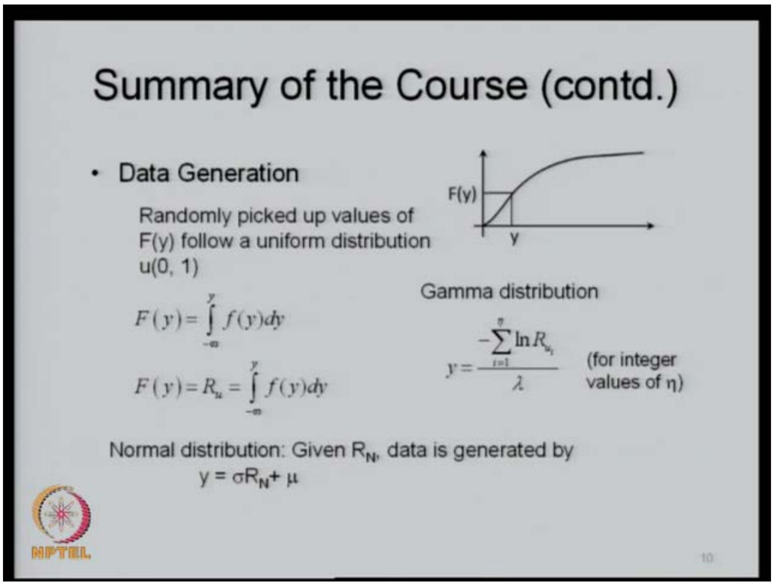


We use this concept of covariance and the joint behavior of the two random variables and then start looking at the sample data, let us say you have a sample data like this these dots of sample data, and then you want to fit a line, straight line then we call it as simple linear regression. This process of fitting a straight line to a observed data set is called as a simple linear regression.

Now, essentially what we do is that, you have a observed point here, y i corresponding to a given x I, then y i cap which is being predicted by the line, so you have a y i cap given by a plus b x i this is the line for a given x i you have a plus b x I, and y i is the observed one, so you have an error y i minus y i cap.

This error in fact, the sum of that error over all the points, if you have n number of points, you sum it over, sum the square of the errors over all the points and that sum of error must be minimized. So, you obtain the coefficients a and b such that, the sum of the squares of these errors is a minimum, and the principal remains the same for even multiple linear regression.

(Refer Slide Time: 24:00)



Then we looked at data generation that means, essentially what we are interested in is that given a distribution, we want to generate several values all of which together will belong to that distribution. Now, this is a important exercise that we did and many applications we have seen, let us say you take a gamma distribution, and you want to generate several values belonging to gamma distribution, where do we apply this?

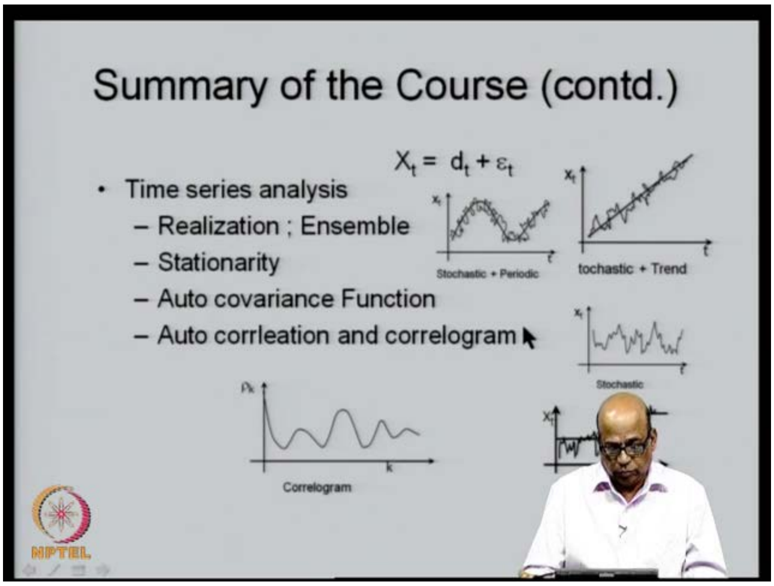
Let us say that you have annual rainfall series and annual rainfall series you have (()) gamma distribution for for the historical data you have satisfied yourself, that the given sequence follows, in fact gamma distribution. Then you would like to generate several such sequences of annual rainfalls all of which belong to gamma distribution, all of which together constitute a all the values in a particular sequence follow gamma distribution, with the same parameters as the historical, as the gamma distribution of the historical data, so that is what we do here.

So, y you generate for integer values of eta, you generate uniformly distributed random numbers as as I mentioned you can use your calculators and get the uniformly distributed random numbers and then generate y. So, keep on changing the uniformly distributed random number here, you get a different value of y, when you generate large number of such values, you get a sequence of random variables, sequence of numbers all of which together will follow random, that is gamma distribution.

Similarly, for normal distribution where we have seen, you generate a normally distributed random number here, you use mu and sigma as obtained from your historical data or as specified and then you generate the data with normal distribution, having those parameters mu and sigma. Now, the basis for this way of generating is that, if you pick up any values on a CDF, randomly you pick up the values on a CDF.

Now, these follow a uniformly uniform distribution, u 0, 1 in the range 1, 0 to 1; now that is a basis on which we generate, we derive several formula for data generation.

(Refer Slide Time: 26:38)



Then we looked at the time series analysis as a as we have seen X t the time series can be written as a deterministic, a sum of deterministic component and a random component. So, the deterministic component can be periodic for example, a periodic component around which the values are fluctuating randomly.

So, you have a stochastic plus a periodic component, you may have a trend like this, so you may have a stochastic component plus a trend, you may have a purely stochastic component, which means there is a no deterministic trend possible at all and you may have a stochastic component plus a jump.

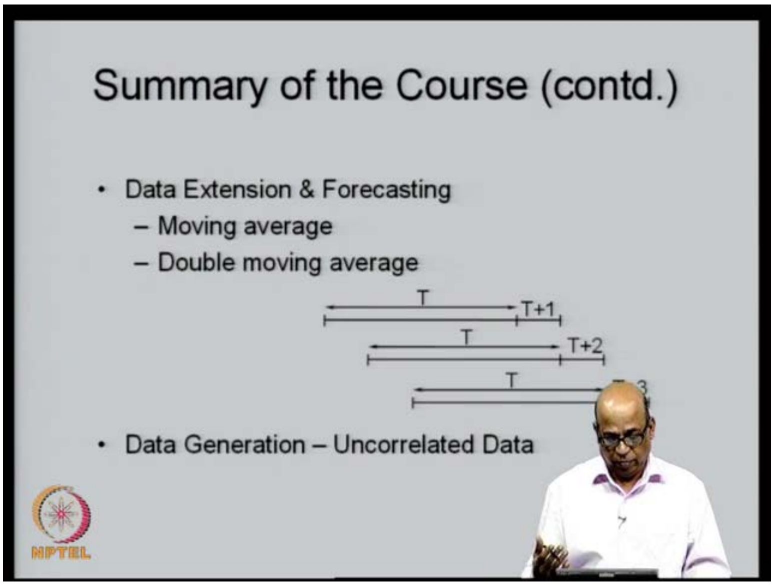
So, it is fluctuating around this point and then there is a sudden jump here, and then it starts fluctuating at a different level. So, you have a stochastic component plus a jump this is stochastic here s t o c h i s t, and then we define the correlogram which is a plot between auto correlation at lag k verses k. So, we have defined auto correlation and auto covariance function much the same way as we had a cross correlation and cross covariance or covariance between two variables starting with that we defined the auto covariance, when you are talking about the covariance of values in the same given time series.

Similarly, auto correlation is correlation of values in the same time series; we also look the concept of stationarity, where we are saying that the probability distributions will remain the same across time windows, across different time windows.

So, we consider the lag tau and then we defined x t plus x t plus tau that is x t and x t plus tau, the distribution of x t and the distribution of x t plus tau will (()) for various values of tau in which case, we call it as a fully stationary time series. Then we defined weekly time series, first order stationarity which is stationary that is the first moment the time series is with (()) to first moment.

Stationary in mean and covariance both, the time series is second order stationary that is second stationary is it is stationary with respect to the first two moments and so on; that is how we define the weak stationarity. Then we also have seen what is a realization and what is an ensemble of realizations and then we looked at the ensemble properties, the time properties, we defined the ergodic process as that particular process where the ensemble properties and the time properties are the same, now these are all important concepts that we have seen in the time series analysis.

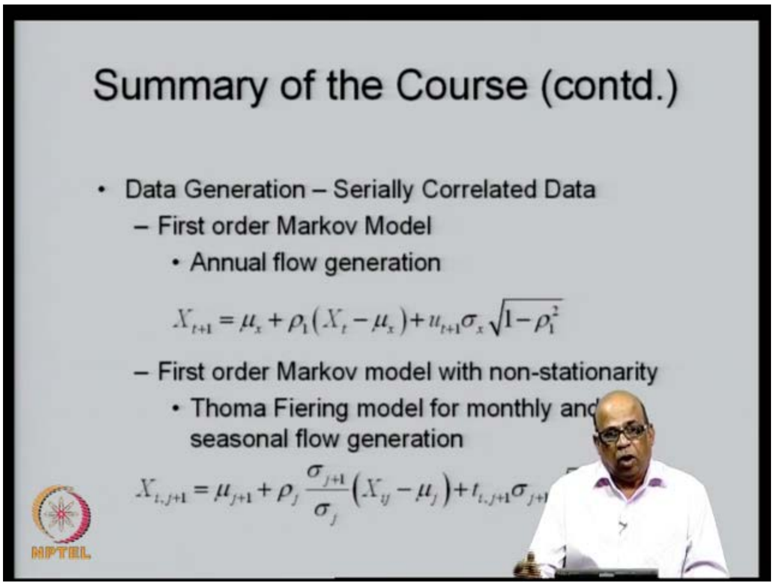
(Refer Slide Time: 29:50)



Then we looked at the essential difference between data extension and forecasting, so in in the moving average for example, we may have a data for time period T and then we are interested in getting T plus 1. So, we get the average of this and then as you move you take out the last value and then you insert this and then keep on taking the averages, that is how we call it as moving average.

Then double moving average on the moving averages you take again an average, so that will define double moving average; we have solved some examples for that. Then we also looked at data generation of uncorrelated data that means totally random data then how to generate the values, using the concept that we have seen earlier for a given distribution.

(Refer Slide Time: 30:44)



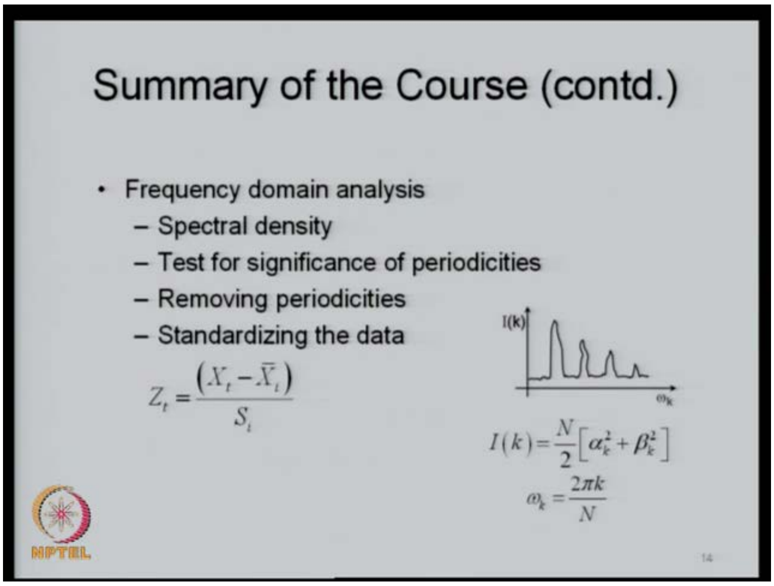
Often in hydrology we get situations where you have serially correlated data for example, you are talking about monthly stream flows, and then you want to generate several values several several values of monthly stream flows following the same distribution as the historical flows have followed, maintaining the same moments; that is you have mu, you have sigma as well as auto correlation of lag 1, lag 1 auto correlation.

Now, these are reproduced by the first order Markov model, recall that we also introduced the name Thomas Fiering model for the first order Markov model. And the first order stationary Markov model, I call it as stationary because, your mu sigma and rho 1 they remain the same, there are only there are only one value associate, there is only one value associated with mu, sigma and rho and that remains constant, so we use this expression to generate annual flow.

Because, it is stationary we use it typically for annual flows, we extended this to a non-stationary first order Markov model by looking at the means and the standard deviations and the lag 1 correlation as a function of that particular time period itself. For example, we may have a mean of June month flow, mean of July month flow and so on.

So, we use the concept that they are different across time periods, and we introduce the non-stationarity in the first order Markov model and this is again a monthly model of Thomas Fiering, monthly version of the Thomas Fiering model.

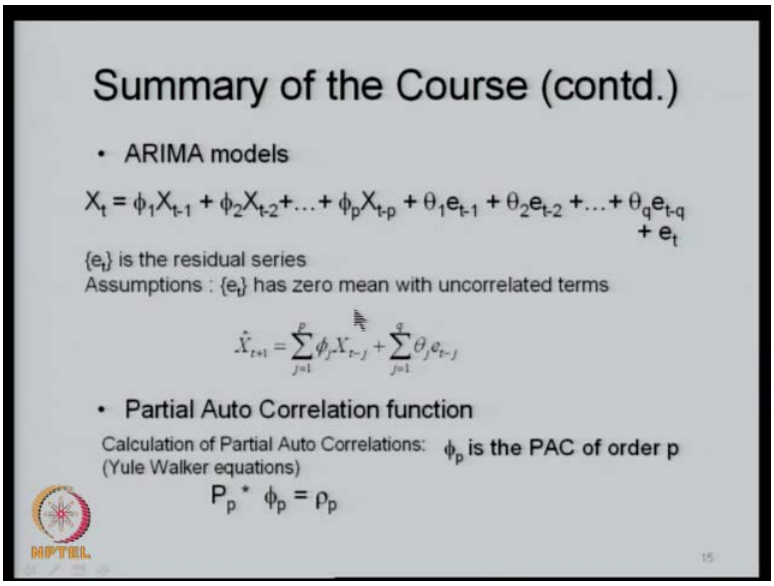
(Refer Slide Time: 32:44)



Then from the time domain, we went to the frequency domain, we looked at the spectral density, we defined the spectral density I k N by 2 alpha k square plus beta k square and then we have also seen, how alpha k and beta k determined from the time series. So, essentially the time series X t on the time (()) was converted into a series of frequencies and then we start looking at the spectra of that particular time series. So, the plot of omega k was verses I k or w k verses I k is called as the spectrum.

Now, once you plot the spectra, the spectrum spectral density you know that there are periodicities in the data, so typically the monthly flow will have certain structure like this of the spectral density, then you have to identify which among these periodicities are in fact, significant. So, we look at the test for significance of periodicities and then we also know by differencing the series you can remove the periodicities and there is other standardization and so on. So, we do certain exercises to remove the periodicities from the data, one of the ways is simply standardizing the data, so by standardizing you remove the periodicities.

(Refer Slide Time: 34:14)



Then we looked at the important auto regressive integrated moving average models, where the interest is in constructing the series X t, we write it as phi 1 X t minus 1 plus phi 2 X t minus 2 etcetera, phi p X t minus p these are the auto regressive parameters.

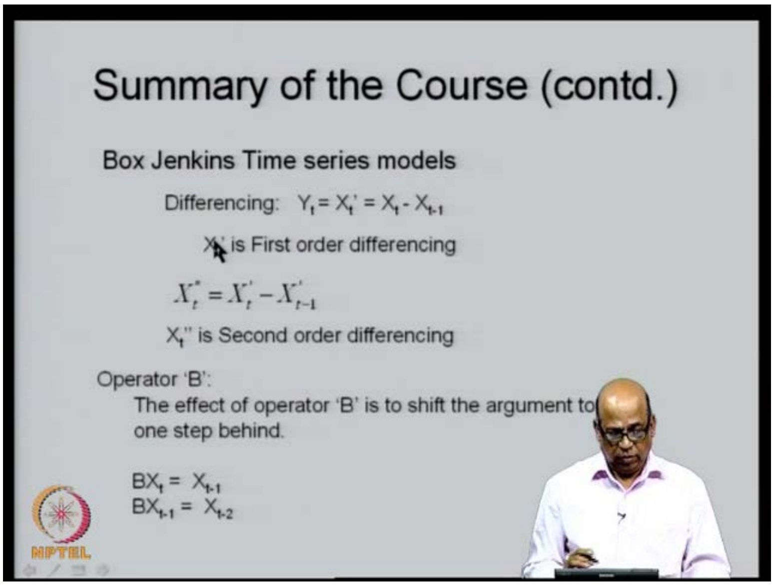
Auto regressive, because X t is dependent on X t minus 1 X t minus 2 etcetera, X t minus p, so there are p terms previous to X t which are included in this model, and they are auto regressive terms, these are the auto regressive parameters, phi 1 phi 2 phi 3 etcetera, up to phi p, then you have the moving average parameters theta 1 theta 2 etcetera, theta q there are q moving average terms, these are defined on the errors plus a random term e t.

So, the e t here is the residual series and this type of models, this particular way of writing is actually an ARIMA model, auto regressive moving average model, but when you write this model on a differenced series it becomes auto regressive integrated moving average model.

So, the term integration here is associated with the differencing for example, if you do the first order differencing, then you will have ARIMA p, 1, q where p is a number of AR parameters, 1 is a order of differencing and q is a number of MA parameters. So, that is how we write ARIMA p d q, d is the order of the differencing, then we introduce the important concept of partial auto correlation, we use the Yule walker equations to get the partial auto correlations.

Partial auto correlations are in fact, an important tool to identify the AR parameters and in fact, we also use for identification of the MA parameters in a particular series, so you when you have a large number of candidate models, you can identify which among those candidate models are better suited, looking at the partial auto correlations.

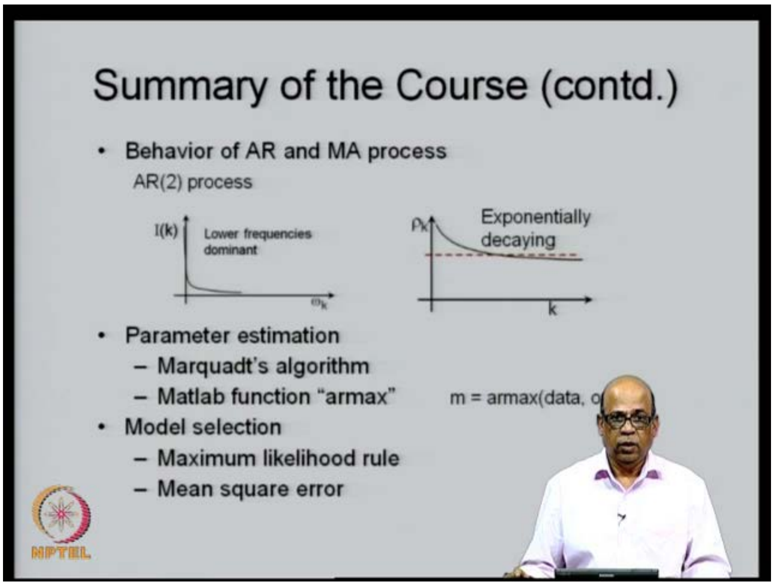
(Refer Slide Time: 36:41)



Then we did a few exercises on the Box Jenkins types of time series models, now these are important models in the sense that many times or (()) applications we straight away use the Box Jenkins type of models. And we looked at the concept of differencing where we defined the first order differencing as X t minus X t minus 1, second order differencing defined on the first order different series X t dash minus X t dash minus 1 and X t dash minus X t minus 1 dash and this is a second order differencing and so on.

And we looked at the operator B, so we are able to express a given ARMA model using the operator B in a compact and elegant form. So, the operator B essentially shifts the particular argument to one time step behind for example, B of X t is equal to X t minus 1 B of X t minus 1 is equal to X t minus 2, that is all. So, the effect of this operator B is to shift the argument to that one step behind, using this we can express several of the models in a compact and elegant model.

(Refer Slide Time: 38:06)



So, we used all of these concepts and looked at how to identify number of AR parameters, number of MA parameters for use in modeling and for example, you look at AR 2 process, theoretical AR 2 process. You will have lower frequencies dominant, this is the spectral density, lower frequencies will be dominant here and then for the same AR 2 process you may have exponentially decaying auto correlations and so on.

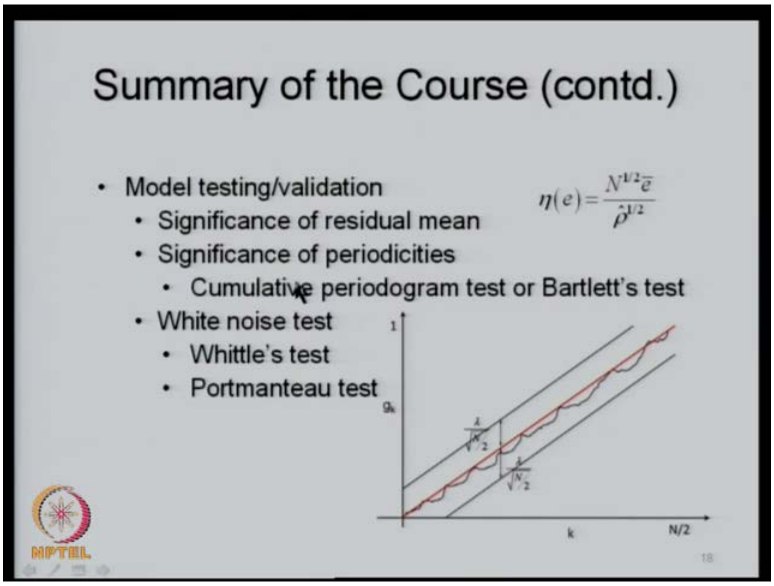
So, given a time series simply, first plot the time series look at the spectral density, look at the correlogram this should give you some idea about the underlying time series that you need to use. And we have seen several such examples, AR 1 MA 1 and AR 2 MA 2 and (()) such processes, it will give you an indication, the moment you will plot all of these, that is a correlogram that (()) density as well as the original time series itself.

You will get an idea of the type of periodicities that are present there and the type of the number of AR parameters that you may have to use and so on. In fact, we looked at the contiguous as well as the non contiguous models, where let (()) have AR 3 process. Now, AR 3 typically may be X t depending on X t minus 1, X t minus 2 and X t minus 3, so all the three previous time period values, time period random variables. However, this is a contiguous models that means, X t dependent on X t minus 1, X t minus 2 and X t minus 3 however, you may have X t dependent on X t minus 1, X t minus 4, X t minus 12.

So, there are three AR parameters however are non (()) and therefore, these are called as a non contiguous models, so non contiguous models are also important in many hydrologic applications. Once you identify the model you look at the parameter estimation, let us say I write AR 1 model as phi 1 X t minus 1, so how do I estimate phi 1 or AR 2 model phi 1 X t minus 1 plus phi 2 X t minus 2 and so on, so you may have p AR parameters q MA parameters to estimate these, we have certain algorithms available and typically we use the Marquadt's algorithm and in the Matlab I have shown, you use the armax function in the Matlab and get the parameters. So, you identify your model (()) data along with your order, that is orders of parameters MA and AR and MA parameters and you will get the parameters.

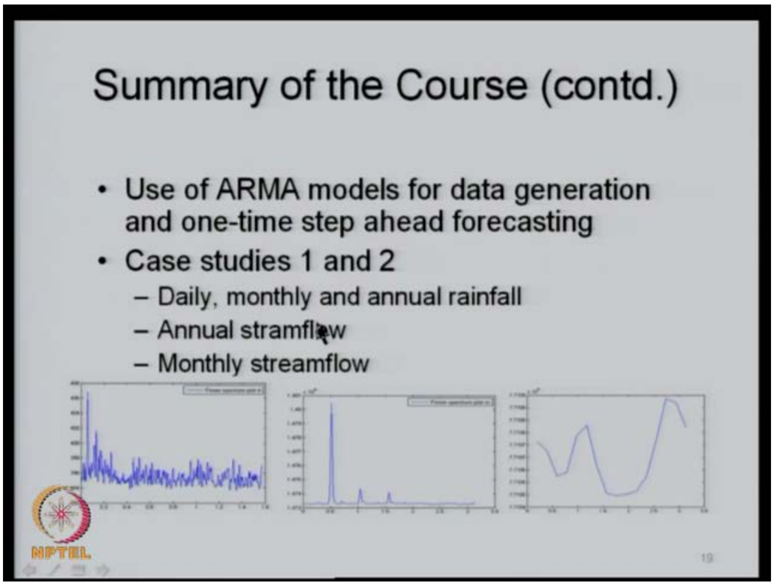
Once you have several of these candidate models, then we look at the maximum likelihood as well as minimum mean square error, maximum likelihood we use for model data generation and minimum mean square error criteria we use for real time forecasting one, one time step ahead forecasting. Once you have the time series model available with you, then you have to look at the validation of model itself.

(Refer Slide Time: 41:30)



So, for the model testing or validation you have several tests that we covered in the course, significance of the residual mean we have to test, significance of periodicities you have to test, typically periodicities are tested with the Bartlett's test or the cumulative periodogram test. You look at this periodogram, if it lies within the significance band, then it is non significant and therefore, the periodicities are absent in the residuals, so these tests you do on the residual series, this is a residual series. Then the series should also be white noise, so the residuals should constitute white noise, so we do Whittle's and portmanteau test to examine that it is in fact, white noise.

(Refer Slide Time: 42:07)

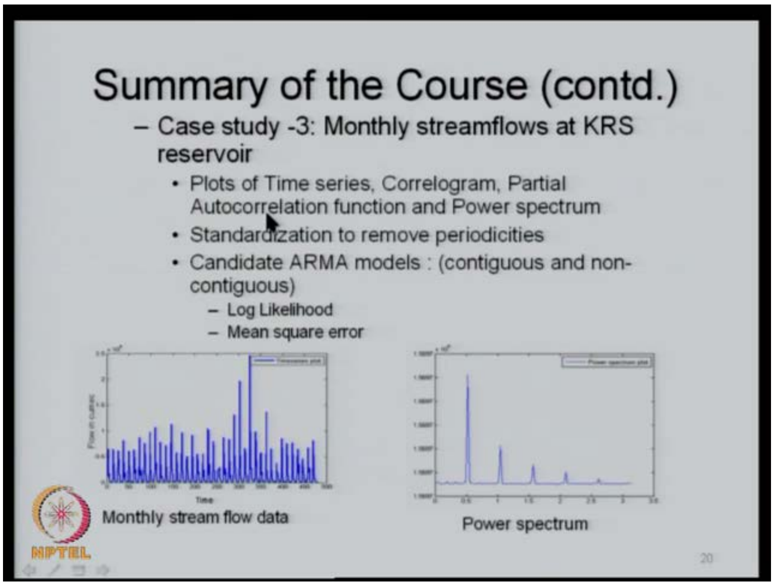


Then in the course, we covered several case studies both for model generation as well as for one time step ahead forecasting, where we looked at the candidate models, then looked at the parameter estimation, from the parameter estimation we went into validation. And then identified from the validation, after the validation you identify those particular models that are best suited for data generation and time that is real time forecasting.

Say for example, this particular example we considered daily, monthly, and annual rainfall, now the power spectrum for daily looks something like this that means there are no significant periodicities it almost looks like a random series. Whereas, if the same series you consider for monthly time series that is you add up the 30 days or 31 days of rainfall and then look at the time series this is how it looks, suddenly the periodicities start coming up.

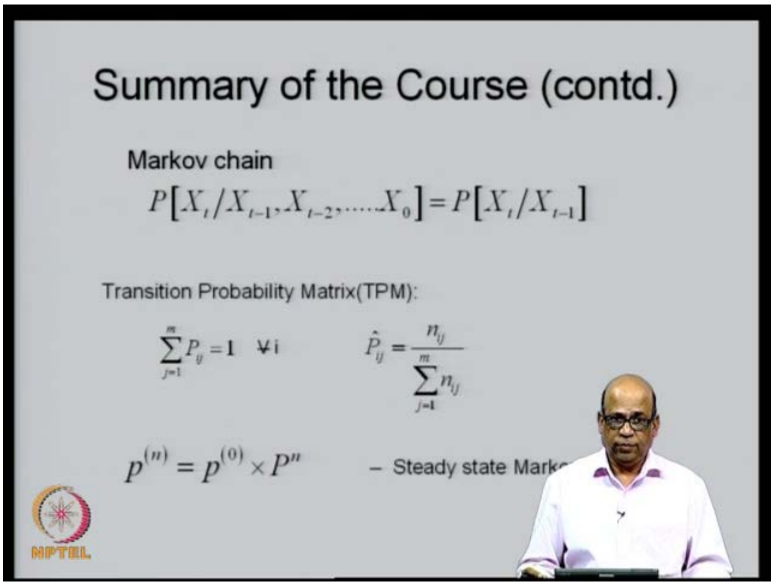
And then the annual thing, annual spectral density the spectral density of the annual rainfall looks something like this, now annual rainfall if you draw the spectral density for (()) distance here, then you may start seeing that it may have certain decadal variations. And in the monthly stream flows you may see that there are periodicities associated with 1 year, may be 6 months, may be 4 months and so on.

(Refer Slide Time: 43:46)



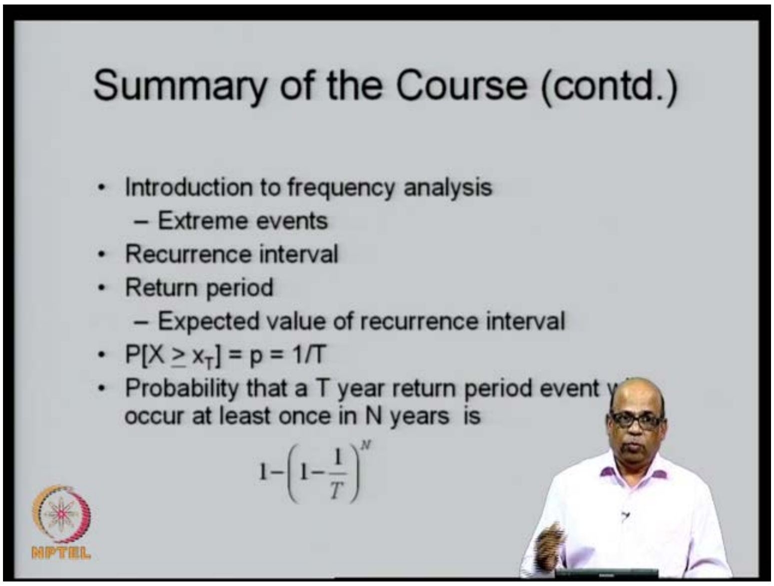
Similarly, we had a monthly stream flow data (()) Kaveri river and then you you have the associated power spectrum, this is a time series, from the time series you get the power spectrum and the power spectrum again shows up periodicities associated with this.

(Refer Slide Time: 44:03)



We went, then to the concept of the Markov chains, so the Markov chain we define probability of X t given X t minus 1, X t minus 2, etcetera up to X naught, that is the probability of the random variable during time period t, given the entire history of the process, X t minus 1, X t minus 2, etcetera X naught. If it can be equated to probability of X t given X t minus 1 that means, the entire memory of this process is contained in X t minus 1, then it is called as a one step Markov chain, associated with this then we talked about transition probabilities, that is P i j t. That is the probability that starting with the state i in period t it goes to the state j in period t plus 1 adjoining the time period that is called as a transition probability. From the transition probabilities we were able to get the steady state probabilities, so steady state probabilities p of n is equal to p of 0 into capital P of n, where capital P of n is a transition probability matrix.

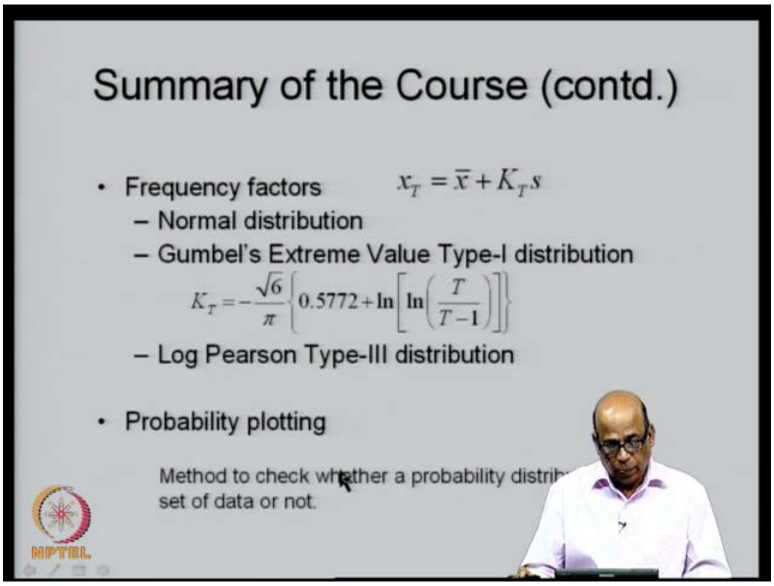
(Refer Slide Time: 45:20)



So, we looked at the extreme events for example, the flood peaks, the minimum flows and so on, which constitute the extreme events and we did the frequency analysis on the extreme events. We defined what is called as a recurrence interval, then from the average recurrence interval, we defined the return period which is the expected value of recurrence interval and in fact, we showed the probability of X being greater than equal to x T.

This capital T is the recurrence interval here is given by 1 over T and we also saw the probability that a T year return period event will occur at least once in N years. Let us say that you are considering 50 year return period and then you are interested in the event that the 50 year return period will occur at least once in next 20 years that is given by this expression.

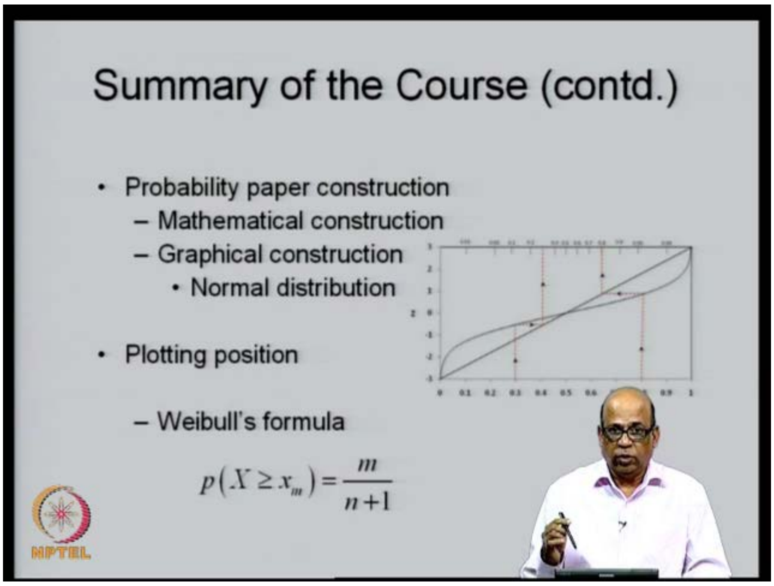
(Refer Slide Time: 46:15)



We use the frequency factors K T to obtain the values associated with a particular time period T, now the K T are the frequency factors, now these are defined for different distributions. And typically we use these as a normal distribution Gumbel's extreme value type I distribution or the Log Pearson type III distribution, we have the K T values tabulated for these distributions for various return periods.

We pick up those K T values and use this expression x bar is the mean obtained from your historical data, this your standard deviation obtained from the historical data; we use the probability plotting to examine whether a given series follows a particular distribution or not.

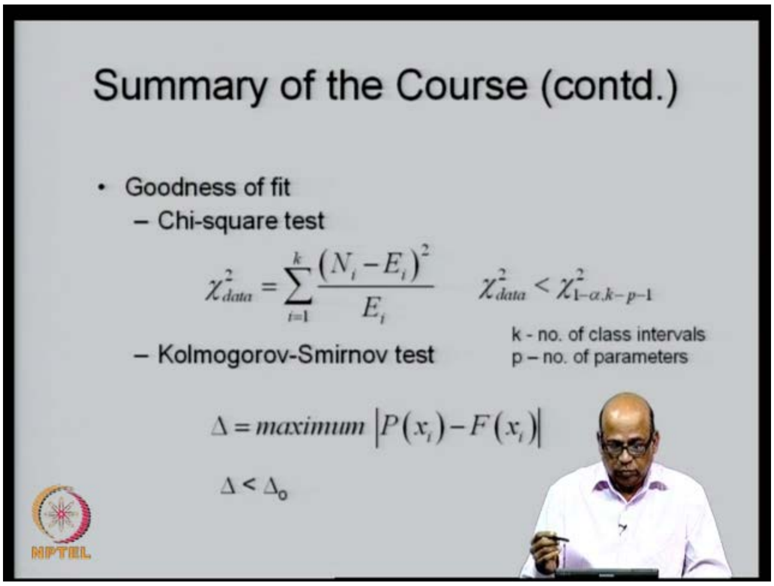
(Refer Slide Time: 47:06)



We use the concept of the probability papers, we have introduced the normal probability paper, the Gumbel's probability paper and so on, and then use a probability plotting positions for example, and the Weibull's plotting position and then plot the data using the plotting position on the probability papers. And if it follows nearly the straight line nearly a straight line on that particular probability paper then it is, it can be assumed to follow that particular distribution.

Now, the graph that I have shown here is the way we construct the normal probability paper graphically. So, recall that this is how your CDF would have looked and then you transfer that such that, it follows a straight line there and then you transform the x axis in this particular case, so that the resulting paper will be a probability paper.

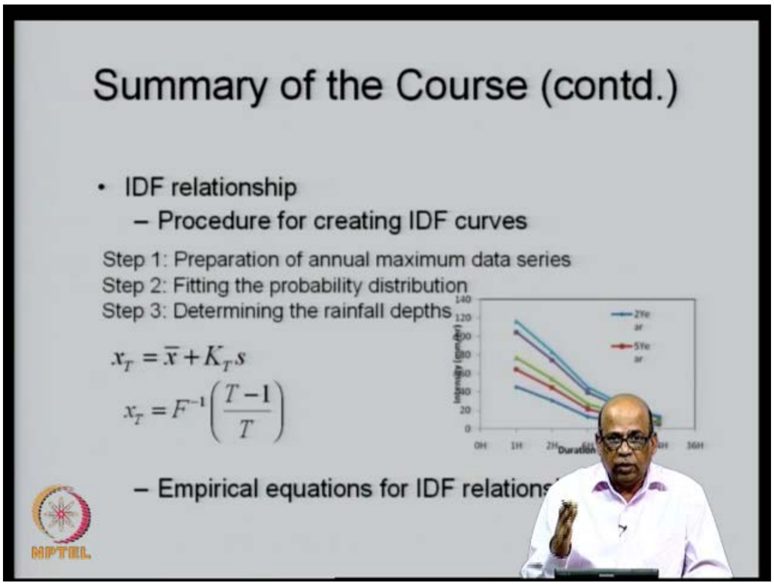
(Refer Slide Time: 47:59)



Then the goodness of the fit that is, let us say that you are suspecting that a particular data set is in fact, normal distribution, then we take check using the chi square test and the Kolmogorov-Smirnov test whether it in fact, follows the given distribution.

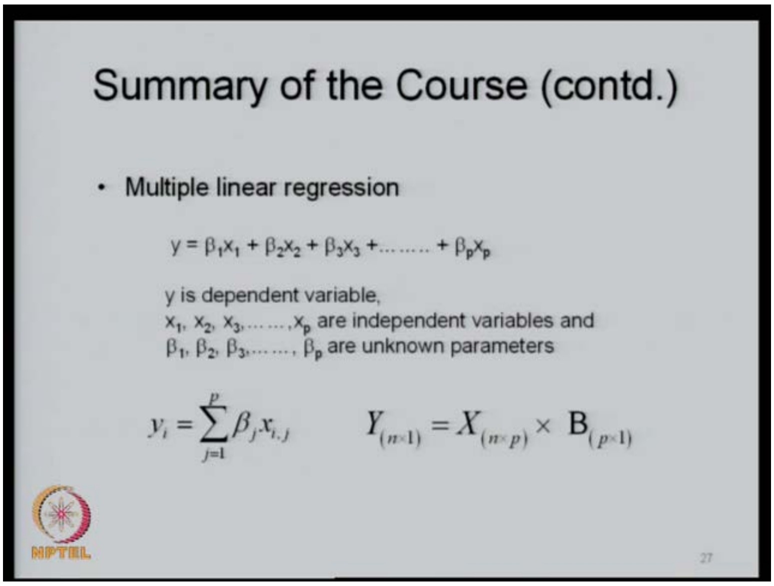
Here for example, you may have a critical value of chi square, where k is the number of class intervals you divide the data into number of class intervals and p as number of parameters and so on. So, these chi square values are given in the tables, you this is arising out of the data and if this is true, then your assumption of that particular distribution is valid. Similarly, in the Kolmogorov test you look at the theoretical CDF values associated a particular i and the probabilities that are coming out of your data.

(Refer Slide Time: 48:56)



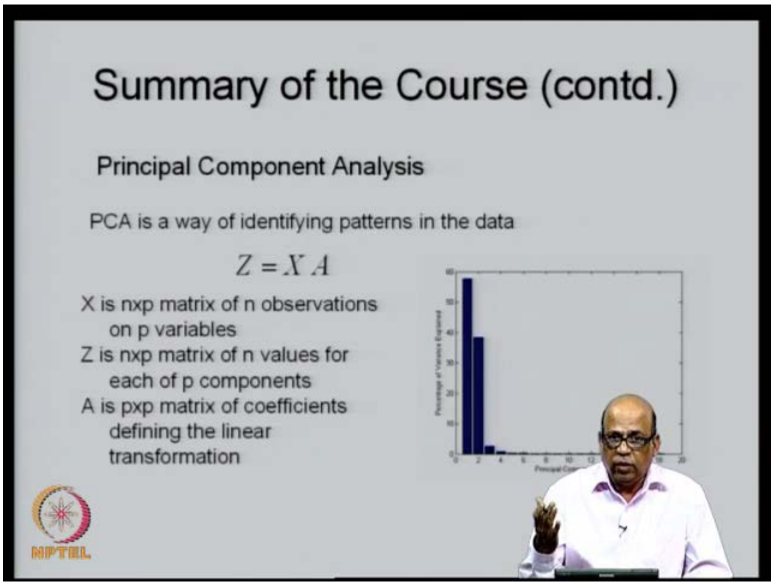
We looked at the IDF relationship that is intensity duration frequency relationship typically we use for flood peaks, to estimate the flood peaks you need the rainfall intensities. And then we use again the frequency factors and obtained x T corresponding to a different particular return period, the duration intensity for several return periods is what we obtain from the IDF relationships.

(Refer Slide Time: 49:26)



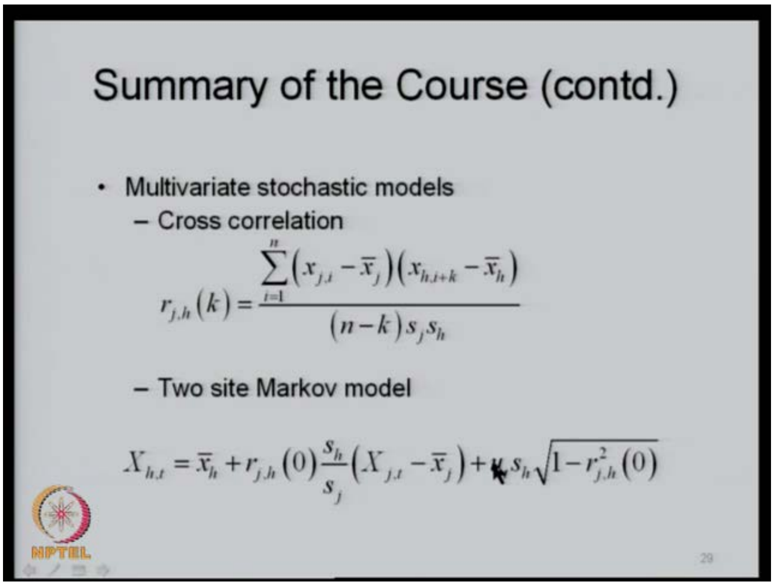
We extend the concept of simple linear regression to multiple linear regressions where your dependent variable is dependent on several such variables x 1, x 2, x 3, etcetera up to x p. And then we had devised (()) following the same principal of minimizing the square errors of obtaining the various coefficients beta 1, beta 2 and so on.

(Refer Slide Time: 49:45)



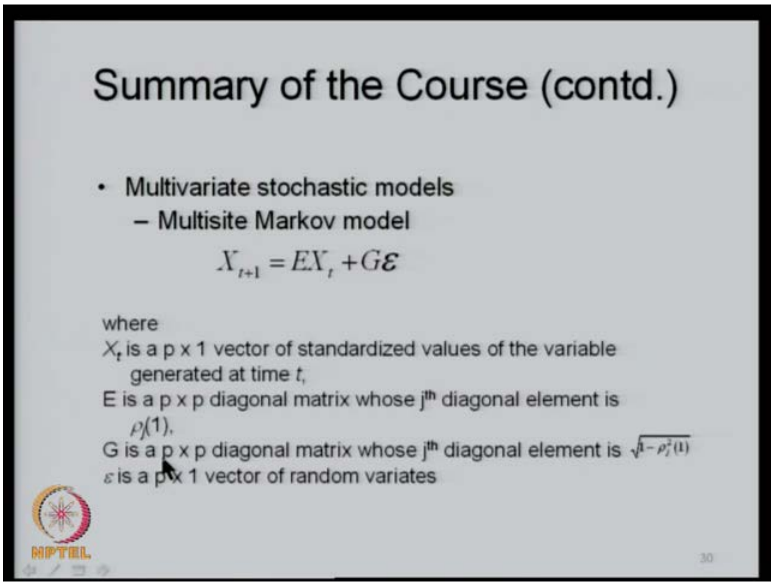
Now, there are cases where there is a multi collinearity that exist that means, these x 1, x 2, x p etcetera, that we consider may be correlated among each other, you have to remove the multi collinearity, you have to reduce the size of the problem in the multiple linear regression and then that is where you adopt the principal component analysis. So, we have seen what we do in the principal component analysis, and then pick up those particular principal components which explain most of the variables most of the variance in the data. And then you used let us say, in this particular case you may use only three principal components, and define the regression on the principal components.

(Refer Slide Time: 50:32)



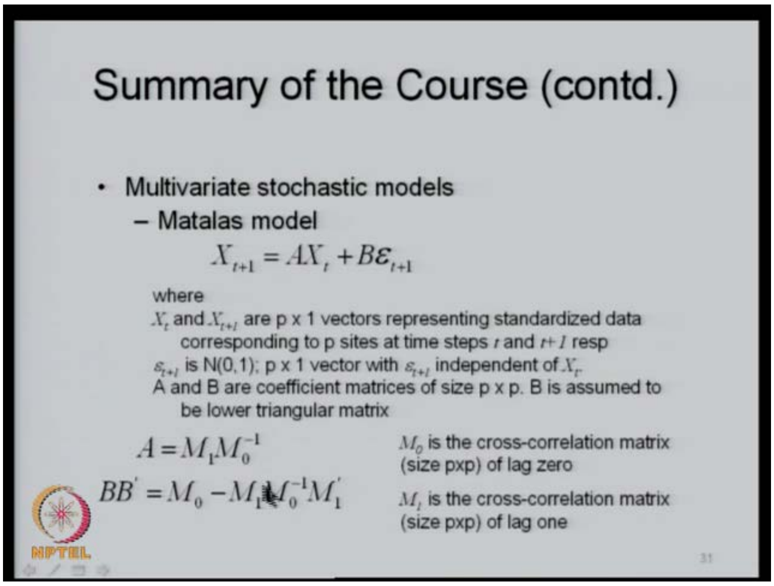
Then we looked at multiple variate stochastic models, where essentially we are interested in the joint behavior this is a station j, station h, and then we are talking about the cross correlation between the data of station j and station h and then we define the two site Markov model essentially for joint behavior of data from two stations.

(Refer Slide Time: 50:58)



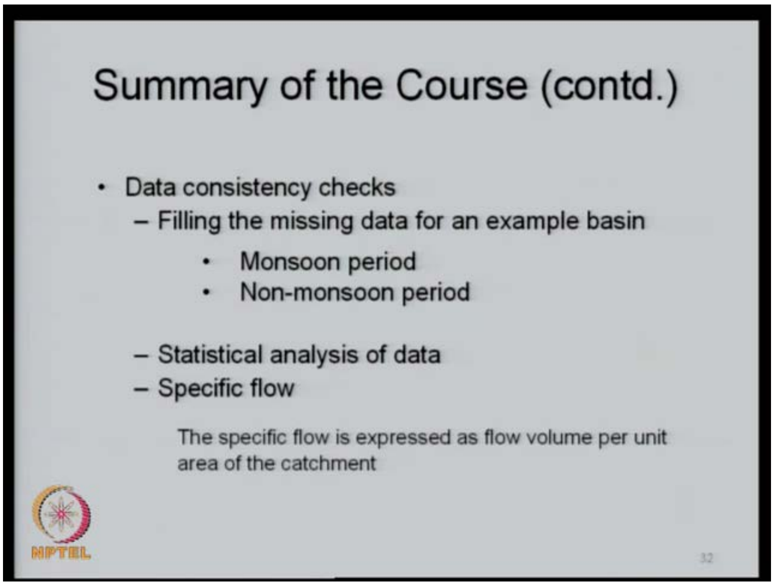
And we derived this multisite Markov model express it as X t plus 1 is equal to E X t plus G E.

(Refer Slide Time: 51:06)



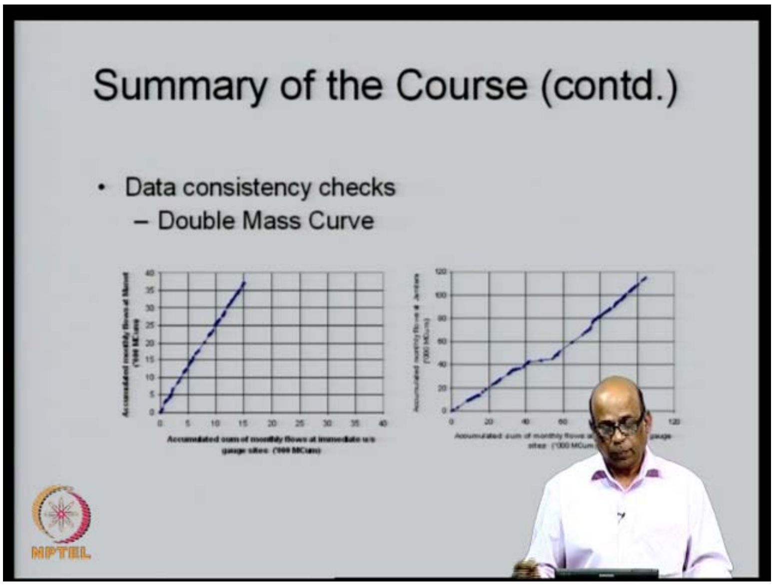
And we have seen how to obtain these coefficients A and B in the Matalas model that is essentially a multivariate stochastic model.

(Refer Slide Time: 51:14)



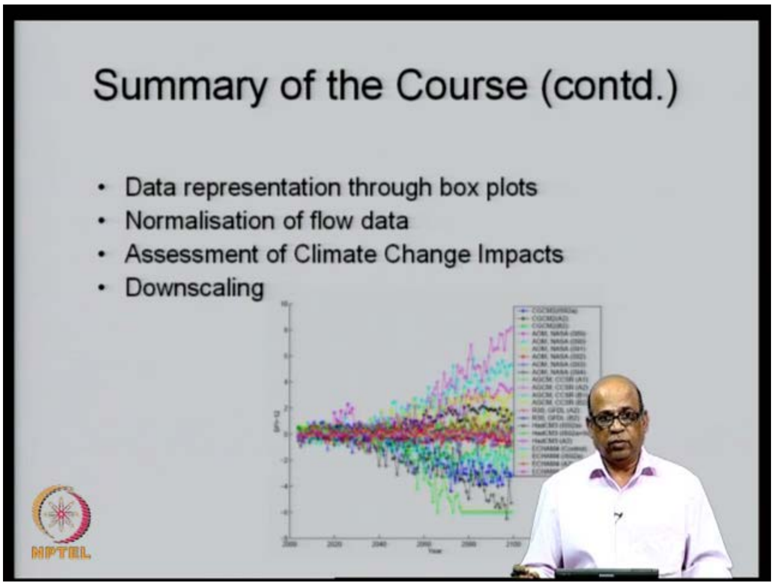
Towards the end we discussed the data consistency essentially using the the mass curve, double mass curve and the concept of the specific flow to look at consistency of flow data specifically at several locations in the (()).

(Refer Slide Time: 51:35)



For example, this is the mass curve; this is another mass curve it shows that there is some break here in this particular period.

(Refer Slide Time: 51:47)



We also looked at the specific flow and how to interpret the data consistency using the specific flow concept. Then we looked at the data representation through the probability that is box plots and how we normalize the flows anthropogenic effects due to reservoirs barrages etcetera, we take it out and then constitute what are called as a naturalized flows.

Then a special lecture was given on assessment of climate change impacts, where we are looking at the future projections of stream flow or in this particular case the SPI standardized precipitation index. As we go into the future the uncertainty bands start increasing, how we address these on certainties using the probability distribution and so on, so that in essence is what we covered in the entire course.

Remember all of these techniques that we have covered in this course, are essentially based on the data that you have, observed data that you have, we fit the probability distributions, we fit the models etcetera, based on the historical data. And then use the same models to project into the future, this is the important concept of stationarity that means, we are saying that the future will be much similar to the past.

But, there is a important concept of non stationarity, the future will not be similar to the past and therefore, you have to account for time varying parameters time varying probability distributions etcetera, into the projections of the future. Now, that is an area of current interest in hydrology where we applied many of these techniques, many of the techniques modified to account for non stationarity and then look at start looking at the non stationarity.

So, this course essentially provided you with some basics of the techniques and tools that we use in hydrologic analysis essentially to address uncertainties. Now, while on the topic of uncertainties you must remember, that the particular way of addressing the uncertainties through the probability theory, that we have covered in this particular course is called as a (()) uncertainty.

In the sense that as information comes in your uncertainty will not change, uncertainty still remains the same, the quantum of uncertainties still remains the same for example, you are looking at the rainfall process, even if you add more and more number of data the type of uncertainty remains the same, you cannot reduce. Because, you are dealing with a natural process whereas, there is another type of uncertainty called as an epistemic uncertainty which will you you can reduce the quantum of uncertainty with more and more knowledge.

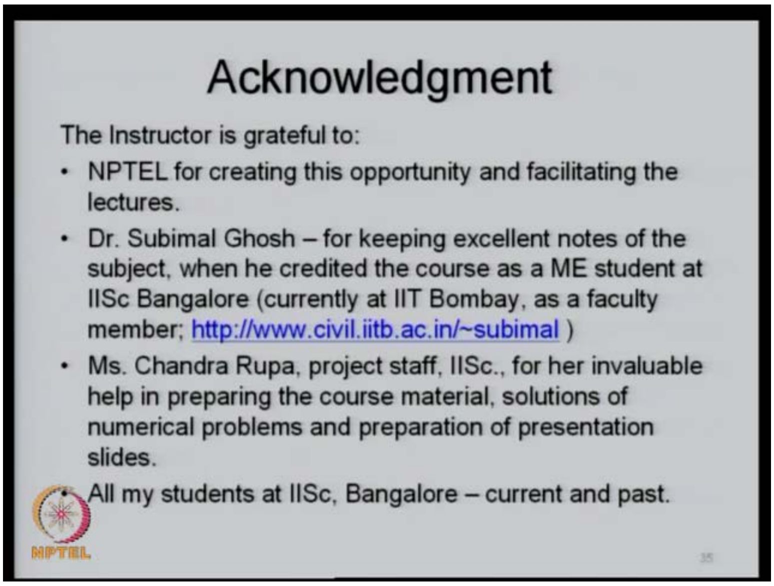
So, the epistemic type of uncertainty arises, because of lack of knowledge and in climate change applications that I just showed in the last class, the epistemic uncertainty is rather important, because most of the uncertainty comes, because of our lack of knowledge.

Whereas, the probabilistic uncertainty is not because of lack of knowledge, because of the inherent type of process itself for example, rainfall it is a natural process and there is a uncertainty associated with natural process and therefore, you cannot reduce that type of uncertainty, you can only address the uncertainty.

So, this you keep in mind I hope you enjoyed the course as much as I did and use these concepts for actual realistic modeling of stochastic processes in hydrology, and make sure that the applications are done with all the limitations of data in mind. As I keep repeating you must never forget about the importance of the data that we have you must be sure of the reliability of the data, you must be sure of the quality of the data, you must be sure of the consistency of the data, you must be sure that these are in fact, the naturalized flow data, that you are using if you are using the analysis on the flows.

So, keeping all these in mind you start applying this technique and then take some more advanced courses on stochastic hydrology and grow well finally, to end this course I would like to thank several people.

(Refer Slide Time: 56:00)



First to NPTEL for creating this opportunity and facilitating the lecture, in particular I would like to thank (()) and Mister Guru Prakash, who have been extremely helpful for me in delivering these lectures.

I would like to thank one of my past students, Subimal Ghosh he is a faculty at IIT Bombay, he took this course along with his other batch mates, when I taught this early on in somewhere around 2004 or 2005 and he kept excellent notes. In fact, I have followed his notes, as I taught in the course and I am extremely thankful to him for keeping excellent notes.

My projects assistant Ms. Chandra Rupa, she has helped me in preparing the course material solutions of numerical problems and preparation of presentation slides I am very thankful to her. And all my students at IISC, Bangalore current and past who have helped me in learning this subject, who have endured my presence in the class room and then gone through this course meticulously.

Most of the case studies that I presented on time series have all come out of the assignments in the short projects that I gave as part of the course to my students. So, with that we end the course, thank you very much for being with me, all the best.