

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

**Lecture No. # 33**  
**Multivariate Stochastic Models- I**

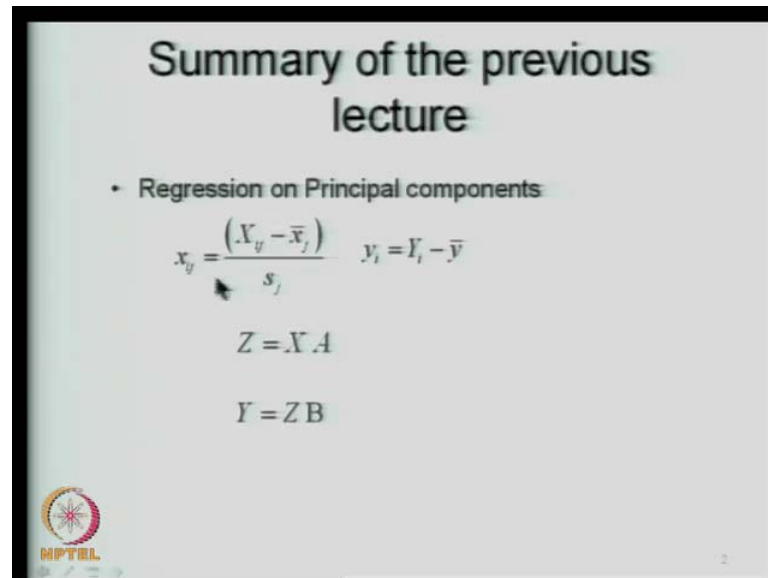
Good morning, and welcome to this the lecture number 33 of the course stochastic hydrology. Over the last few lectures, we have been discussing about the principal component analysis, multiple regression and so on, where essentially we want to develop a relationship between the dependent variable  $y$  and several independent variables  $x_1, x_2, \text{etcetera}, x_p$ . And in the previous lecture, in the last lecture, we discuss the principal component analysis, which has the advantage that it will reduce the size of a problem also, it will account for the correlations among the independent variables. And what do I mean by accounting for, let us say  $x_1, x_2, x_3 \text{ etcetera}$  they are mutually correlated, and that leads to what is called as multicollinearity in a multiple linear regression and therefore, you need to have variables, which are not correlated with each other.

By carrying out the principal component analysis, this is what you achieve. The original set of  $x_1, x_2, \text{etcetera}$  up to  $x_p$ , the  $p$  independent variables you convert into a new set of components  $p_1, p_2, p_3, \text{ etcetera}, p_p$  components you derive and these new components we call that these components are linear transformation of the original set of variables, and these are uncorrelated. And we also know how much of the variance in the dependent variable is explained by each of these principal components, and based on that we can decide how many of principal components we would like to retain in the regression.

So, let say you had 10 variables, and then out of 10 variables, you may just so, 10 variables will lead to 10 principal components essentially 10 eigen vectors, and then out of these 10 variables, 10 principal components you may want to retain only 6 or only 3 as we discussed in the previous lecture. So, principal component analysis achieves two things for us one is, that the size of the problem can be reduced instead of dealing with 10 variables each of the 50 years of data and so on, you may now deal with only 3 variables. So, the size of the problem is reduced, and the correlations among the

independent variables that you had original  $x_1, x_2, \dots, x_p$ , the 10 variables that we consider in the example those correlations are removed when we do the transformation. So, this is what we discussed in the previous class.

(Refer Slide Time: 03:05)



The slide is titled "Summary of the previous lecture" and contains the following content:

- Regression on Principal components:
$$x_j = \frac{(X_j - \bar{x}_j)}{s_j} \quad y_i = Y_i - \bar{y}$$
$$Z = X A$$
$$Y = Z B$$

In the bottom left corner, there is a logo for "CAPTEL" with a star-like symbol. A small number "2" is visible in the bottom right corner of the slide.

I will just go through that what we did, so that original variables then we use the principal components on the regression analysis so, instead of doing the regression on the original observed variables we do the regression on the principal components. So, the original variables are first standardize by standardize you now know that, we deduct the mean and divide by the standard deviation, and that is how we get the transformed variables. So, this defines the vector  $X$ , the matrix  $X$  where you have  $p$  variables and then  $n$  values associated with each of them.

So, the size of  $X$  is  $n$  by  $p$ ,  $n$  variables,  $n$  observation along the  $n$  rows and then  $p$  stations along the you will have  $p$  columns associated one associated with each of the variable. So, this is  $n$  by  $p$  now this is the matrix consisting of the eigen vectors or the principal components so, you will have the size  $p$  by  $p$ . So, you should have  $p$  stations remember you get the principal components based on the covariance matrix, and the covariance matrix will be of the size  $p$  by  $p$ , and you will get the matrix  $A$  which is of size  $p$  by  $p$  and that is how you get this is  $n$  by  $p$  and this is  $p$  by  $p$ . So, you will get this as  $n$  by  $p$ .

Now, this is the transform data from the original data using the principal components you transform the data into  $Z$ , we now develop the regression relationship between the

dependent variable  $Y$ , and the transformed data  $Z$  by using the parameters  $B$ .  $Y$  is dependent variable so, this is a single variable, it has  $n$  observations and therefore, it is a vector  $n$  by  $1$ , and this is  $n$  by  $p$ , and this is  $p$  by  $1$ . You have  $B$  as a vector of parameters  $\beta_1, \beta_2, \dots, \beta_p$  there are  $p$  terms; so  $\beta_p$ . So, this is  $n$  by  $p$  this is  $p$  by  $1$ , and that is how you get  $Y$  as  $n$  by  $1$ .

So, this is how we regress and then we have also seen the expression for  $B$  which is essentially the same as what we did in our multiple linear regression, where we regress the original variables  $X$  directly not even transformation, we used directly the  $X$  variables and the regress  $Y$  with  $X$ . So, this is what we covered in the previous lecture. So, we do the principal component analysis choose those principal components which explain most of the variance, and then based on those principal components we do the regression, this is what we call as regression on principal components.

We also saw a example of 10 variables that means, there were 10 rainfall stations and then you are estimating the yield or the run off at a particular location as a function of all these 10 variables, rainfall in 10 stations in the watershed, and in that particular example if you recall what we saw is that if you put all the 10 variables in that original form, that means,  $y$  regress upon  $x_1, x_2, \dots, x_{10}$ , you got certain  $r$  square value which is a essentially the major of goodness of the regression relationship.

However the  $x_1, x_2, \dots$  etcetera maybe correlated, so, you would like to remove the correlations, and fit another regression relationship based on the principal components. So, which we did we used 6 principal components, and when we did the principal component analysis we saw that 95 percent of the variance is explain by the 6 principal components, and the develops the regression relationship. Then we reduce the number of principal components to 3, where 85 percent of the variance was expand which means essentially as we reduce the number of components you are sacrificing some information content.

And when you reduce it to 3 you could explain only 85 percent of the variance, but the interesting point that is when you reduce the 3 reduce it to 3 principal components, the coefficients of the first 3 principal components remain the same as that was used earlier. Now, this is the interesting feature of regression using principal components this we have seen in the previous lecture. Now we will take this analysis further, this discussion

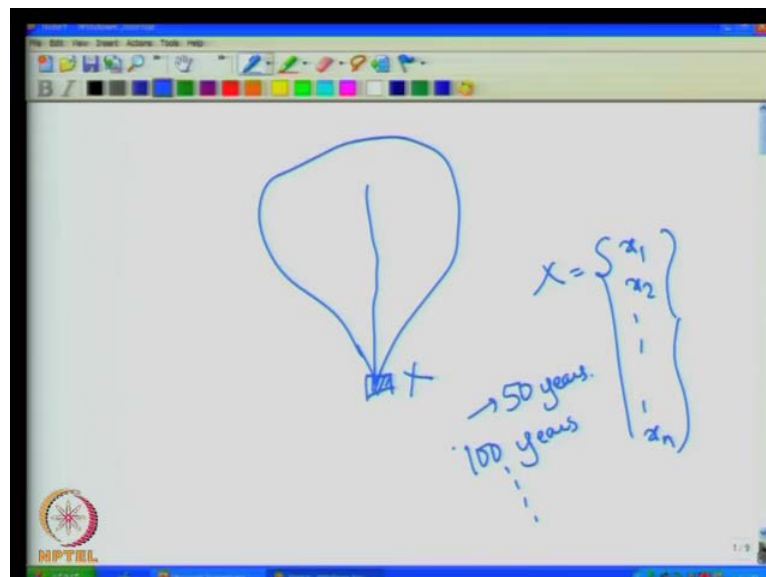
further and see where we get situations in hydrology where you have to deal with simultaneous behavior of 2 or 3, 2 or more variables and you are interested in generation of data on 2 or 3 or more variables and that is what leads us to multivariate stochastic models.

(Refer Slide Time: 08:33)



So, in today's lecture we will essentially focus on development of multivariate stochastic models, let us have look at some situations where we encounter these problems in hydrology.

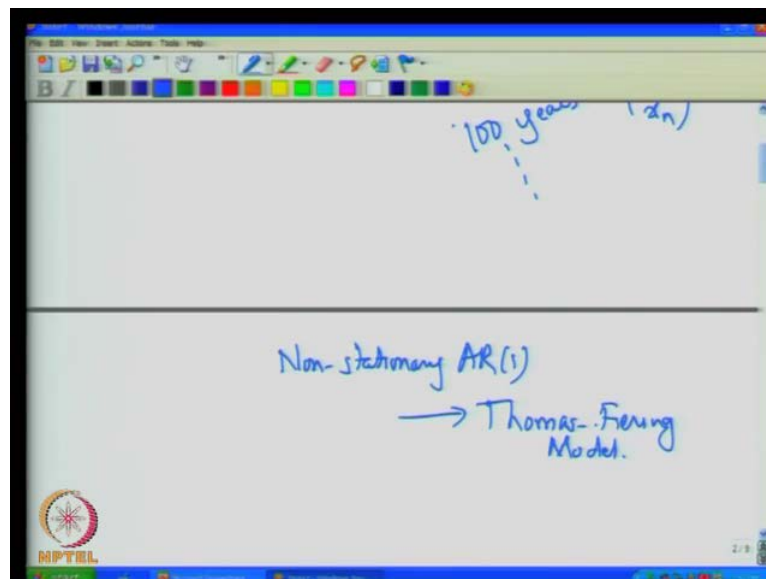
(Refer Slide Time: 08:50)



Let say that you have a watershed like this, and then there is a stream, you have a stream gage here at this location; you have observed data at this point let say this is the random variable  $X$ , the stream flow at this location we define it as random variable  $X$ , and this  $X$  you have data, observed data like  $x_1, x_2, \dots, x_n$  there are  $n$  years of data available if it is a annual stream flow we are talking about 100 years of data is available. We have seen earlier how to generate data on  $X$ , let say you had 50 years of data, and then you want to generate another 100 years of data, and many such sequences not one sequence of 100 years, but several such sequences of 100 years which you can use for making any decision at this location.

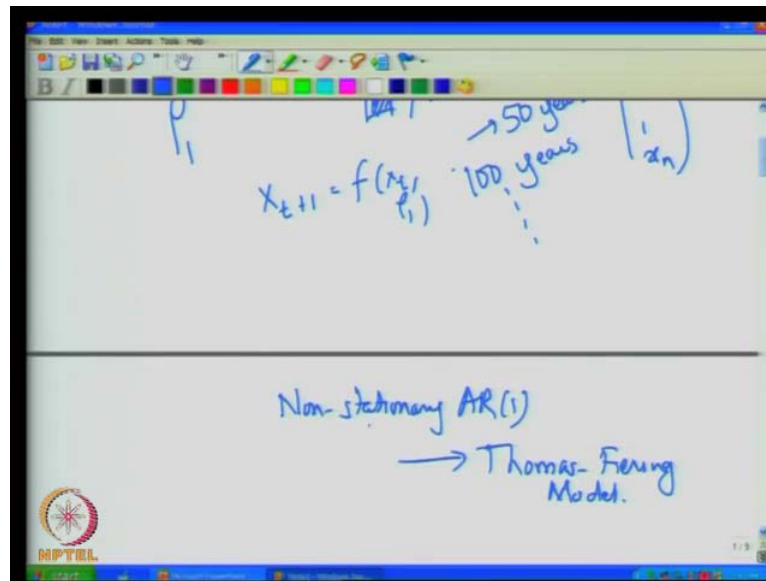
Let say you want to build a reservoir or some such thing, you will we will based your decision not on one sequence of observed values, but on several sequences of generated values using the observed historical data. We have seen several methods on this, but we will focus on one of the methods which is, if you recall in one of the earlier lectures we talk about the Markov chain model.

(Refer Slide Time: 10:30)



So, we said it is a non-stationary AR (1) model, auto regressive model with of order 1 this is what we talk, and then we said it is called popularly known as the Thomas Fiering model. So, what does Thomas Fiering model do? It considers the auto correlations, and specifically the lag 1 auto correlation of the variable.

(Refer Slide Time: 11:20)



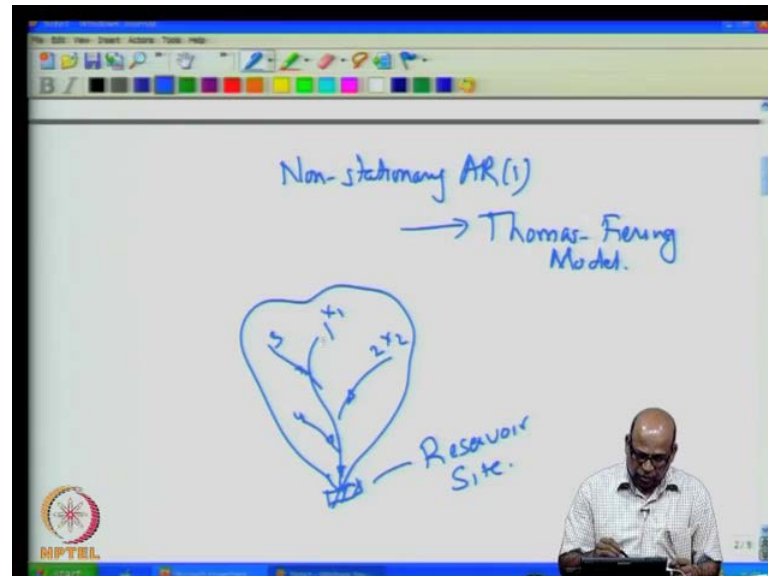
That is, when you have a random variable  $X$ , you define  $\rho_1$  as the lag one auto correlation this is the time series now. On this time series you get the lag one auto correlation and then use these lag one auto correlations to generate  $X_{t+1}$  value has a function of  $X_t$ , and lag one auto correlations, and we have seen the expressions for this now the model that we discussed earlier and called as Thomas Fiering model, which is essentially a non-stationary auto regressive model of order 1.

It preserves the mean, the standard deviation, and the lag one auto correlations of the data what do I mean by preserve? when you generate this data let say, you generate the data for 100 years, next 100 years the mean of that generated data will be close to the mean of a historical data, observed data. Similarly, the standard deviation with close to the historical observed data, and the lag one correlation is close to the observed data, and we call it as a non-stationary model when we introduced the variation of the mean from time period to time period, and especially for monthly flows, we use this kind of models monthly stream flows.

We use the kind of model that I discussed earlier where the mean of june month is different from the mean of july month, different from mean of august month, exedra. So, these non-stationarity in the mean, as well as non-stationarity in the standard deviation, and the lag one correlations. The lag one correlation between the august and july is much different from the lag one correlation between march and february for example. So, these

differences in the mean, standard deviations, and the lag one correlations they are all incorporated into the Thomas Fiering model.

(Refer Slide Time: 13:22)



Now look at a situation where you are not dealing with one stream flow, but you have the same watershed and then you are interested in building a reservoir, let say at this location you want to build a reservoir site, and the flow that are coming here are sums of 2 or more streams flows. Let say you have a stream flow coming here from this point **from this point from this point**, and the main stream flow; so, you have 4 stream flows. 1, 2, 3, 4 and the actual flow at this location is sum of 1, 2, 3, and 4 and you want to generate the data for construction of for the making decisions on that is our capacity here which means that you would like to generate data on 1, 2, 3, and 4 together.

Now if they are all independent let say this is the random variable  $X_1$ , this is the random variable  $X_2$  and so on, all of these are independent or can be assume to be independent or you can afford to ignore the dependence among themselves then what do we do? we can consider each of them as a single time series, and generate exactly the way we did earlier, that means, this becomes over time series by itself and then you generate it using the single site models, remember the earlier model that we talk about is a single site model, because we are not considering any other influence on this time series.

However, in most hydrologic situations when we are talking about such problems where 2 or 3 streams are together contributing to a particular location, there will be a significant

correlation among  $X_1$ ,  $X_2$ ,  $X_3$ ,  $X_4$  and so on, and therefore, it is important when you are doing the data generation where slightly moving away from our multiple linear regression, and then we are now looking at building stochastic models for data generation and data forecasting. Do not get mixed up with what we just discussed in the previous lecture. In previous lecture our focus, was on developing a relationship between the flow at this location and rainfall at several locations.

The rainfall was causing the flow, and that is why we build the regression relationship between the flow and the rainfall at various locations. The problem that I am discussing now is much different from what we have discussed in the multiple linear regression, here we are talking about data generation at this location, that is statistical data generation at a particular location when it depends on when it is contributed, the flow is contributed by several streams like this, and these flows are all correlated. So, you cannot afford to ignore the correlations among them self, among each of these variables when you are considering the flow generation at this location.

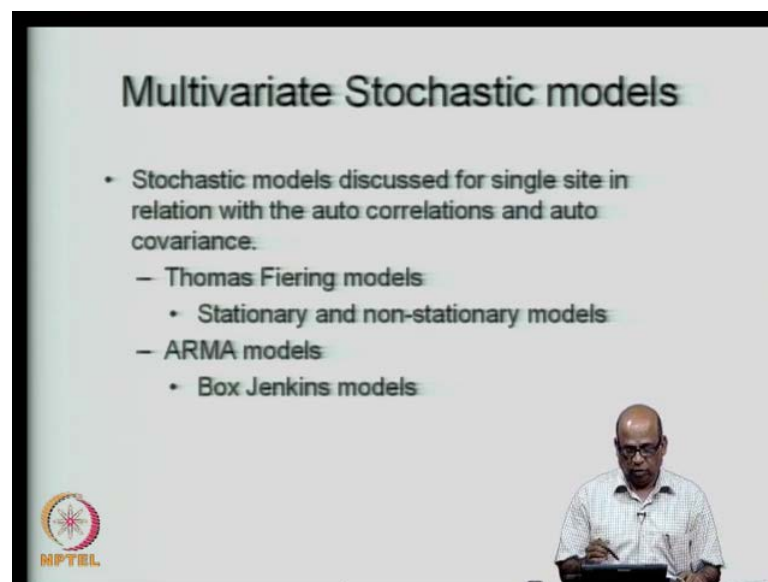
In the Thomas Fiering model, that we discussed earlier what did we do? We took the serial correlation or the auto correlation. You were talking about only one time series and then you were talking about the auto correlations. Whereas, when we come to problem such as these where you are talking about simultaneous behavior of several random variables  $X_1$ ,  $X_2$ ,  $X_3$ ,  $X_4$ , etcetera and then you want to incorporate these features of the simultaneous behavior into the data generation model or into the stochastic models. Then, we need to look at not only the auto correlations that exist in the series  $X_1$  and in the series  $X_2$ , but also the cross correlations between  $X_1$  and  $X_2$ .

So, your model should be capable of capturing the relationship between  $X_1$  and  $X_2$ , and between the values of  $X_1$  itself with respect to time. So,  $X_1$  has a time series,  $X_2$  has a time series, etcetera. So, it has its own lag one correlations though should be captured as well as the dependence on other random variables here  $X_1$ ,  $X_2$ , etcetera. This is what we will do in today's lecture. How do we incorporate the information content arising out of several such random variables? and the specifically we will talk about the flows, because these are kinds of problems that we frequently get in hydrology where there are many streams and then you want to generate data at a particular location looking at cross correlation of these.



One thing that, we must I must repeat is that if there are situations, where these dependences can be ignored or your reasonably sure that these are all independent or the cross correlations are negligible. In such situations, you can use the single site models generate data on  $X_1$ , data on  $X_2$ ,  $X_3$ ,  $X_4$ , etcetera and then add them up to get the data at this location. This is what you would have done if the cross correlations were, in fact, negligible. However when we cannot ignore the cross correlations, we must have a mechanism or we must be models, where the cross correlations as well as the auto correlations are taken into account. Let see how we do this? So, that leads us to the topic of multivariate stochastic models, and this is what we will discuss today.

(Refer Slide Time: 19:30)



So, the stochastic models that we discussed earlier for single site especially the Thomas Fiering type of models, you also has stationary models as well as non-stationary models. As I just explained when you are dealing with annual flows for example, they can be stationary models and when you are when you come to non-stationary models you are typically dealing with the monthly flows or 10 daily flows, 15 day flows, etcetera where there is a significant difference or significant variations in the means from one type period to another time period, and that non-stationarity you are incorporating in the models.

Then we also have discussed earlier general family of models auto regressive moving average models of course, the Thomas fiering model, that I discussed is one specific case

of the ARMA model auto regressive moving auto regressive of order 1, and then in the ARMA models we had talk about Box Jenkins models where you expressed  $X_t$  as, let say  $\theta_1$  into  $X_{t-1}$  plus  $\theta_2$  into  $X_{t-2}$  etcetera plus  $\theta_1$  into  $\epsilon_{t-1}$  like that you have auto regressive terms, as well as the moving average terms just go through the previous lectures and revise this.

These were all single site models in the sense that we were talking about only one time series there. So,  $X_t$  was a one time series. So, I may put this as I am sorry we can write these as  $X_t$  as a time series. So, with respect to time you have measured the values on the random variable  $X$  and that is what constitutes the time series. So, we were talking about single site models. So, at one site you have the observed data, and then you are building the models for  $X_{t+1}$  given  $X_t$  we are building models on  $X_{t+1}$ , and this is what we use for long term generation of the data as well as for forecasting of the data.

(Refer Slide Time: 21:42)

**Multivariate Stochastic models**

First order Markov process:

$$X_{t+1} = \underbrace{\mu_x + \rho_1 (X_t - \mu_x)}_{\text{Deterministic component}} + \underbrace{\epsilon_{t+1}}_{\text{Random component}}$$

$\epsilon \sim \text{Mean 0 and variance } \sigma_\epsilon^2$

$$X_{t+1} = \mu_x + \rho_1 (X_t - \mu_x) + u_{t+1} \sigma_x \sqrt{1 - \rho_1^2}$$

And just to complete our discussion on the previous coverage of part. We wrote the first order Markov process as  $X_{t+1}$  is equal to  $\mu_x$  look at these we had one time series and from  $X_t$  you are generating  $X_{t+1}$  here. So, this is a stationary model, because we are treating the  $\mu_x$ , which is the mean as constant so, it is not changing with respect to  $t$ . So, this is constant and then you have  $\rho_1$  which is lag one auto correlation and

this is the random component. So, you have the deterministic component and the random component.

Now, when we used epsilon as mean 0, and variance epsilon sigma e square, we wrote the problem in this fashion that is,  $X_{t+1}$  is equal to  $\mu_x$  which is the mean of the random variable  $X$ , this is lag one correlation, and this is a random term, and this term here make sure that you get the same variance as the variance in your  $X$ , and this is your standard deviation of  $X$ . When we use this kind of model as I mentioned, the generated data will preserve the mean, the standard deviation, and the lag one correlation, and these are very popularly used in hydrology.

(Refer Slide Time: 23:12)

**Multivariate Stochastic models**

First order Markov model with non-stationarity, for stream flow generation:

$$X_{i,j+1} = \mu_{j+1} + \rho_j \frac{\sigma_{j+1}}{\sigma_j} (X_{i,j} - \mu_j) + \epsilon_{i,j+1} \sigma_{j+1} \sqrt{1 - \rho_j^2}$$

$\rho_j$  is serial correlation between flows of  $j^{\text{th}}$  month and  $j+1^{\text{th}}$  month.

$\epsilon_{i,j+1} \sim N(0, 1)$

When we introduce the non-stationarity, what we did? We make the  $\mu$  which is the mean change from time period to time period. So, you are generating for time period  $j$  plus 1, and then you will consider the mean of flows during that period we generally apply these models for stream flows so, I keep on using flows, but any random variable these models are applicable, and this is a  $\rho_j$  which is the serial correlation between  $j$ -th month and  $j$  plus 1-th month. What we are doing? We are generating for the  $j$  plus 1-th month using the information on the  $j$ -th month and therefore, you need the correlation between  $j$ -th month, and the  $j$  plus 1-th month.

Let say july you want to generate base on june data and therefore, you need the correlation between the flows in june month, to the flows in july month, and that is how

we generate this and these are the random components and typically we use the normal standard normal variates which is  $N(0,1)$ . All of these we have discussed earlier, but this is what leads to our discussion today and that is why I had to revise this. These are single site models so, as long as your concern only of flows at a particular location, you can use any of the single site models that you have discussed earlier.

(Refer Slide Time: 24:43)



**ARIMA Models**

ARMA (p, q)

$$X_t = \underbrace{\phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p}}_{\text{AR of order 'p'}} + \underbrace{\theta_1 e_{t-1} + \theta_2 e_{t-2} + \dots + \theta_q e_{t-q} + e_t}_{\text{Residuals of order 'q'}}$$

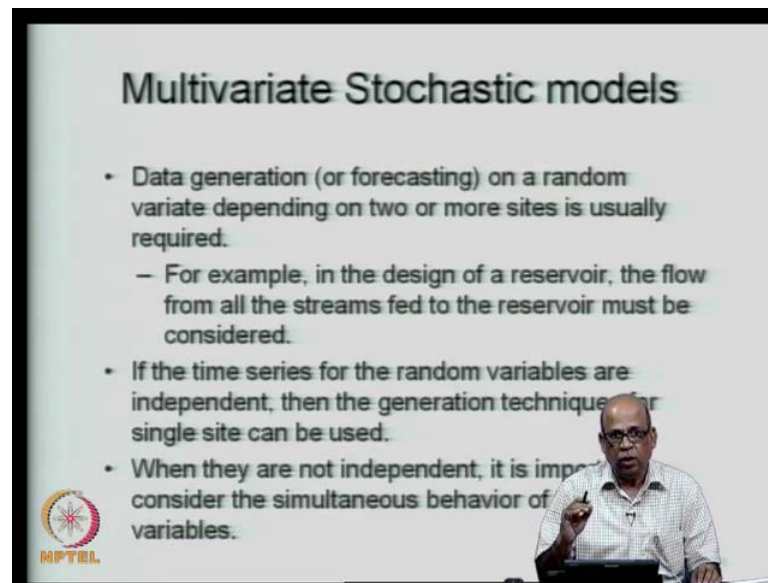
$\{e_t\}$  is the residual series  
Assumptions:  $\{e_t\}$  has zero mean with uncorrelated terms

$$\hat{X}_{t+1} = \sum_{j=1}^p \phi_j X_{t-j} + \sum_{j=1}^q \theta_j e_{t-j}$$

Similarly, in ARIMA models you have AR of p parameters p order that is, phi 1, phi 2, up to phi p, and then you have residuals or the MA parameters of q, and then you have a nice term and, this is how we write it in compact form. Now, all of this is passed now. So, let us move forward these are all single site models where the concern is on data generation, specifically data generation these models are used for data generation at single site.

(Refer Slide Time: 25:19)



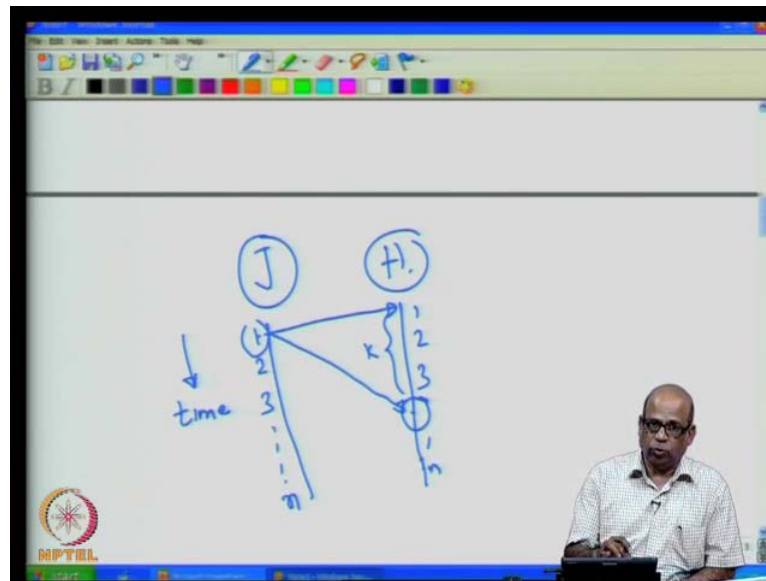
### Multivariate Stochastic models

- Data generation (or forecasting) on a random variate depending on two or more sites is usually required.
  - For example, in the design of a reservoir, the flow from all the streams fed to the reservoir must be considered.
- If the time series for the random variables are independent, then the generation technique for single site can be used.
- When they are not independent, it is important to consider the simultaneous behavior of these variables.

Now in the multivariate case we are interested in data generation of more than one sites. As I said you are looking at the design of a reservoir at a particular location, now the flows may have been contributed by several of these random variables  $X_1, X_2, X_n$  and then therefore, and therefore, you are interested in generating data at all these locations simultaneously, that means, you have to take into account the dependence of one dependence of flows in one site on the flows in another site, and that is how you need to do the data generation, these are called as multisite data generation models in hydrology, and the product of classes multivariate stochastic models.

Now as I said if you can assume that these several of these random variables that we are talking about are all independent then you use the single site models. However in many situations it so, happens that we cannot afford to ignore the cross correlations among these different random variables.

(Refer Slide Time: 27:04)



So, we now define let say we are talking about a situation where we will examine, so that we do not miss the points of definitions correctly. Let say you have sites J and H, these are two different sites, on the site J you have observations. So, this is time, let say I have observation 1, 2, 3, exedra, n number of observations. Similarly, on site H, I have observations 1, 2, 3, etcetera up to n and these are the time steps  $t$  is equal to 1,  $t$  is equal to 2, etcetera the time series J on J, that is  $X_J$  if you call  $X_J$  has its own serial correlations, that is  $\rho_k$  for various  $k$ s lag is equal to 1, lag is equal to 2, exedra, its has its own auto correlations.

Similarly, the site H has its own autocorrelation. I use autocorrelation and serial correlation analogously. So, both of them mean the same. In addition they will have a cross correlation between this time period and the same time period on H, that is lag zero cross correlation. They will also have a lag  $k$  cross-correlation. Let say this is  $k$  times step apart. So, the correlation between the variable on J at time period  $t$  with the variable on H at  $k$  time steps apart is what is called as lag  $k$  cross-correlation between a variable on J and variable on H which we denote as  $X_J$  and  $X_H$ .

So,  $X_J$  and  $X_H$  have also have cross correlations and these cross correlations can be at various lags, what do I mean by that? the flow in J during time period 1 or the other way round we take flow in H during a particular time period may be influence by flow on J during  $k$  time steps behind. Let say you are august here and then this august flow on the

flow in H, August flow in site H may be influenced by let say August, July, June. June flow on site number J, and these dependences are captured by the cross correlations at a specified lag.

In general when we talk about cross correlations lag zero is what is implied. So, we are talking about this as cross correlation. Cross correlation between X J and X H is simply at the same time we take and then we compute the cross correlations, but we are now introducing a concept of cross correlations at lag k, and when k is equal to 0 this is what we mean that is the simply cross correlation. So, we use this concept now and introduce in the models the cross correlation at lag k when we are generating the data.

So, that is what we mean correlation of a random variable between two sites is cross correlation, this is what we have seen then lag zero cross-correlation is the cross correlation at two points on the same time period. Let say actually the cross correlation, that means, you have two sites and your taking the same time period then you are computed the cross correlations that is lag zero.

Then lag k cross-correlation which is denoted as  $r_{j, h}(k)$ , j is the station, h is the station. So, we are talking about flow in site j and flow in site h, and k is the lag it is a time lag. So, lag k cross-correlation is the correlation between random variable at site j flow at site j you understand it as flow at site j with the random variable at site h, which is the flow at site h with lag time k, and as I mentioned lag time is you have time series here on site j and site h you have time series you lag it by k time period and then compute the cross correlation that is what is called as the lag k cross-correlation.

(Refer Slide Time: 32:02)

The slide is titled "Multivariate Stochastic models". It contains the following text and formula:

$$r_{j,h}(k) = \frac{\sum_{i=1}^n (x_{j,i} - \bar{x}_j)(x_{h,i+k} - \bar{x}_h)}{(n-k)s_j s_h}$$

where

- $n$  is the total number of pairs of observations on  $X_j$  and  $X_{k+}$
- $x_{j,i}$  is the  $i^{\text{th}}$  observation on  $X_j$
- $\bar{x}_j, s_j$  are the mean and standard deviation of observations on  $X_j$

In the bottom right corner of the slide, there is a small circular logo with the text "MPTEL" and a photograph of a man in a white shirt and glasses, who appears to be the presenter.

And we use the same expression as we use for cross correlation except that, we now lag the series by  $k$  time steps. So, we are talking about the lag  $k$  cross-correlation between the flows in site  $j$  and site  $h$ , this is given by you take the flow in site  $j$  for the time period  $i$ ,  $i$  is equal to 1 to  $n$ ,  $n$  number of observations are available, and deduct the mean of the site  $j$  mean of the flows at site  $j$ . Similarly,  $x_{h,i+k}$  this indicates at site  $h$ , but now will record the time period  $k$  time step ahead. As I shown you have the site  $j$  here and site  $h$  here and then what we are talking about is if you are at this location this  $h$   $i$  plus  $k$  will be here. So, this is  $i$ , this is  $i$  plus  $k$ , and this is the  $k$  time steps ahead.

So, that is what we do here on the site  $h$  you will lag it by  $k$  time steps and then take this and this is the mean of flows in  $h$ , this is standard deviation of the flows of site  $j$ , standard deviation of flows at site  $h$  and, because we are using lag  $k$  this becomes  $n$  minus  $k$ . Please, also refer to our earlier discussions on auto correlation etcetera where we do the lagging that is what we are talking about lag  $k$  auto-correlation etcetera is much similar to that except that, we are now talking about cross correlations and those lags. So, this is how we compute the lag  $k$  cross-correlations.

Now this is an important input into the multivariate stochastic models in the single variate stochastic models of the single site models the lag  $k$  auto-correlations were important and specifically the type of models that we saw lag one autocorrelation were important. Whereas in the multivariate stochastic models not only the lag  $k$  auto-



correlations on each of these sites, but also the lag k auto, lag k cross-correlation between the site j and the site h between the flows at site j and the flows at site h they are important.

(Refer Slide Time: 34:19)



**Example – 1**

Obtain the lag one cross correlation of annual rainfall data in mm at two sites A and B.

Year	1	2	3	4	5	6	7	8	9	10
Annual rainfall at site A (mm)	5496	7797	7392	7061	6564	5919	5053	3951	4280	5910
Annual rainfall at site B (mm)	5713	6934	6275	6641	6675	5605	5144	5116	4722	6869

Year	11	12	13	14	15	16	17	18
Annual rainfall at site A (mm)	5145	6384	5679	6021	6733	8151	4151	4200
Annual rainfall at site B (mm)	5226	7313	6068	5876	6044	8384	5149	

So, let us see how we compute this now? We have two simultaneous sites simultaneous observations on two sites, let say this is site A, and this is site B, and both of them belong to same hydrologic region or homogeneous hydrologic region, and we want to build cross correlations among these two.

Now, these type of problems are also important in applications where we are talking about prediction in ungaged basins, you may want to use the little data that is available by ungage we mean where the data is small or data is completely absent, you may want to use the data that is available in a nearby stream build a correlation structure build the dependent structure and then use that dependent structure to generate the data at the location where data is negligibly small. So, we have for 19 years now annual rainfall at site A and annual rainfall at site B now these are the data that are available. So, using these we will now look at how to develop the cross correlations at various lags.

(Refer Slide Time: 36:10)

### Example – 1 (Contd.)

Site	A	B
Mean	5926	6069
Std.dev.	1250.1	914.9

lag one cross correlation of sites A and B is given by

$$r_{A,B}(1) = \frac{\sum_{i=1}^n (x_{A,i} - \bar{x}_A)(x_{B,i+1} - \bar{x}_B)}{(n-1)s_A s_B}$$



12

So, we are now talking about lag one cross correlations. So, we will see how to get the lag one cross correlation same thing can be done same procedure can be adopted for obtaining different lags. Now, the mean values which are again in a depth units millimeters at site A is this and at site B is this, now the lag one cross correlation at site sites A and B using this expression now, this is A and B and lag one correlation. So, k is equal to 1 is what I will put here. So, I write  $r_{A,B}$  that is the lag one cross correlation between the flows at site A and site B, and this is lag one and this is the data on A and this is the data on B one time step ahead.

So, if you are looking at i-th time step you take the i plus 1-th time step on the data on B, and this is the mean of flows at B, mean of flows at A which are obtain here from the data and this is standard deviation at A and standard deviation at B. So, this is what we use and compute the lag one cross correlation between the flows at site A and the flows at site B.

(Refer Slide Time: 37:22)

S.No. (i)	Annual rainfall at A	Annual rainfall at B	$(x_{A,i} - \bar{x}_A)$	$(x_{B,i+1} - \bar{x}_B)$	$(x_{A,i} - \bar{x}_A) \cdot (x_{B,i+1} - \bar{x}_B)$
1	5496	5713	-430		
2	7797	6934	1871	865	-371836
3	7392	6275	1466	206	385557
4	7061	6641	1135	572	838719.5
5	6564	6675	638	606	687965.4
6	5919	5605	-7	-464	-296072
7	5053	5144	-873	-925	6328.587
8	3951	5116	-1975	-953	831772.6
9	4280	4722	-1646	-1347	2660008
10	5910	6869	-16	800	-1316760
11	5145	5226	-781	-843	13354.06
12	6384	7313	458	1244	-971409
13	5679	6068	-247	-0.95	234.44
14	6021	5876	95	-193	-18333
15	6733	6044	807	-25	-20175
16	8151	8384	2225	2315	-5150875
17	4151	5149	-1775	-920	1633750
18	4200	5359	-1726	-710	1231420
19	6704	6197		128	

So, this is typically done using any of a spread sheet programs typically used Microsoft excel or any spread sheet program that you are use to, this is the data on A, annual rainfall at A, and annual rainfall at B and then you have computed already the mean of A and mean of B. So, this is the data rainfall at both these locations I am sorry I indicated it as flows, but these are in fact, the rainfall at two locations just make that correction these are not the flows you are looking at a watershed here, and then there is there are two stations A and B, and we are interested in getting the cross correlations between the rainfall at A and rainfall at B.

Even if there was flows the procedure remains the same, but for this particular problem I am taking the rainfall at these two locations. So, this is the record available on the rainfall at A and this is rainfall at B we have computed  $\bar{x}_A$ . So, simply take out  $\bar{x}_A$  from this. So, you constitute this series, because we are talking about lag one we will take the second data for this, this is  $x_{A,i} - \bar{x}_A$ . So, this is corresponding to this value that is 5496 minus the mean of that is this. Then 6934 the second value we take for site B, because when i is equal to 1, I am talking about the second value here and that is what leads to 865 then,  $x_{A,i} - \bar{x}_A$ , this is what I have to multiply  $x_{A,i} - \bar{x}_A$  into  $x_{B,i+1} - \bar{x}_B$ .

So, from these I multiply to this and then get this particular value, similarly this multiply by this I get this value like this we compute this term, and then we submit and get the

sum of all these values and that is what gives you this term, this summation is what we sum along this and that we use. So, this is sum comes out to be this number 3373079 that is a sum of this. Again just understand that I will take i is equal to 1 and for i is equal to 1, I will use for rainfall at A the same value and for rainfall at B, I use the second value, because we are talking about lag one value here. Let say you are talking about lag 4 value now for this value I will take i is equal to 1 and next one I will take i is equal to 5.

(Refer Slide Time: 40:13)

**Example - 1 (Contd.)**

$$\sum_{i=1}^n (x_{A,i} - \bar{x}_A)(x_{B,i+1} - \bar{x}_B) = 3373079$$

$$r_{A,B}(1) = \frac{\sum_{i=1}^n (x_{A,i} - \bar{x}_A)(x_{B,i+1} - \bar{x}_B)}{(n-1)s_A s_B}$$

$$= \frac{3373079}{(19-1) \times 1250.1 \times 914.9}$$

$$= 0.164$$

So, you are taking essentially i plus k. So, this is how you tabulate and then obtain the sum of that and that sum is here in this particular case it is 3373079 for lag one and then you use these formula n minus 1, n is 19. So, and s A is the standard deviation at site A, and standard deviation at site B, rainfall at site B and then you get the lag one cross correlation as 0.164. So, this is how you get lag k cross-correlation for varying k. Let say this was lag zero, that means, lag zero cross-correlation it simply the cross correlation between the two variables in which case you can see from the formula that this will be again x Bi itself minus x bar B. So, this is just the cross correlation between the two variables. So, we used typically the r A,B are in terms of your population notations it will be rho A,B at various lags in our multivariate stochastic models.


(Refer Slide Time: 41:25)

## Multivariate Stochastic models

Multisite Markov model (Two sites):

- Model preserves mean, variance, skewness, lag one serial correlation and lag zero cross-correlation (Haan 1977).
- One site is to be selected as key site.
- Selection may be based on the length of the data and the quality of the record.
- Consider  $j$  as the key site and  $h$  as the subordinate site to key site  $j$ .
- A sequence of observations is generated using single site generation technique.

Ref.: Haan, C.T. (1977) Statistical methods in Hydrology, Iowa State



Now we will start with multisite Markov model I discussed at the beginning of this lecture the Thomas Fiering model. Thomas Fiering model is the single site is a single site Markov model and it is a single site one step, Markov model in as much as we consider  $\rho$  one that is lag one auto correlation. Now we use the single site Markov model and then start developing multisite Markov model, by adding the cross correlations now in the single site Markov model we had only the cross correlations and we start with that and then start developing the cross correlations into that particular model.

The single site Markov model or the Thomas Fiering model that we talked about earlier preserves the mean, the standard deviation, and the lag one correlations and therefore, the multisite Markov model that we want to develop should preserve the mean at both the locations, let say you are talking about the two locations it should preserve the mean at both the locations, it should preserve the standard deviation or the variance of both the locations, it should preserve the lag one correlations at both the locations, and it should also preserve the cross correlation, because there is a cross correlation among these two.

So, starting with the single site model we now proceed to multisite models where not only the mean, standard deviation, and lag one correlations of a particular site are preserved, but also the cross correlations which indicate, in fact, the dependence of one site on the dependence of the variable on one site and the variable on the other site these are also preserve. How do we do this? We will start with a case of only two sites first, and then we will generalize it, let say you have a watershed and you are interested in getting the generation on two sites and these two sites are correlated the flow the random

variables on these two sites typically will take the flow, the stream flows that these two sites are correlated and therefore, we cannot assume them to be independent and therefore, we cannot use the single site models for both of this.

When you have two sites that is the data on two sites we will choose one of the sites to be a key site. Now this procedure is explain in Haan I am using the same procedure. We will choose one of the sites as the key site, remember you want a data generation on both the sites, and one of the sites has longer data and may be the quality of the data is also better. We use this data we use this particular station as the key site, and generate using the single site model we generate the data on the key site.

Let say I denote the key site as j, and the other site we call it as subordinate site. So, there is a key site j and the other one is the subordinate site to key site j. On the key site we will generate using the single site models. So, using single site generation technique, we generate the data on the key site.

(Refer Slide Time: 45:08)

**Multivariate Stochastic models**

- A cross-correlation model is used to generate values of site h based on generated values at site j.


$$Y_{h,t} = \bar{x}_h + r_{j,h}(0) \frac{s_h}{s_j} (Y_{j,t} - \bar{x}_j) + u_{t,h} s_h \sqrt{1 - r_{j,h}^2(0)}$$

*j* and *h* refer to two sites, in this model

**First order Markov model with non-stationarity (single site)**

$$X_{i,j+1} = \mu_{j+1} + \rho_j \frac{\sigma_{j+1}}{\sigma_j} (X_{i,j} - \mu_j) + \epsilon_{i,j+1} \sigma_{j+1} \sqrt{1 - \rho_j^2}$$

*i* is year, *j* is month in this model

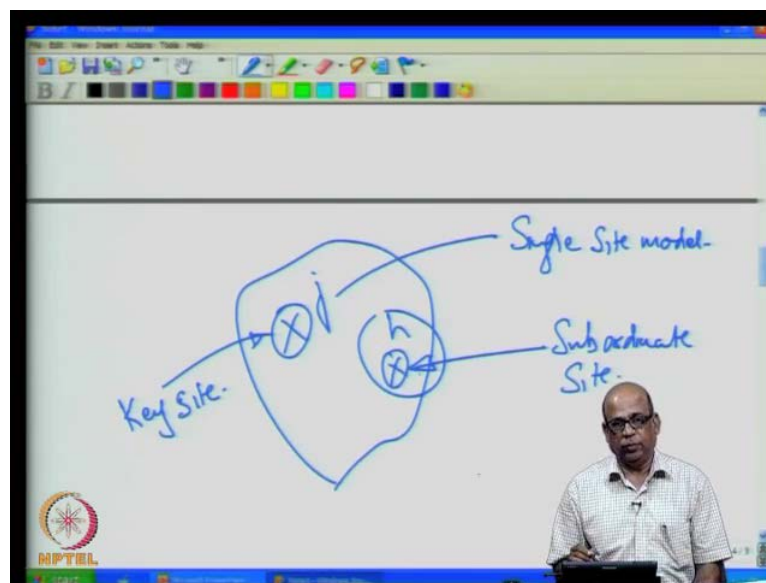


How do I do that? this is your first order Markov model for a single site this is non-stationary model. So, I will use this model with non-stationarity, now in this particular way of writing it i is the year and j is the month using the same. So, we are generating it for several years now. So, this is year and this is month, j plus 1 is a month, and this is our Thomas Fiering model that we have used and these are the standard normal deviates. We use this particular model either in stationary form or non-stationary form. We will

start with the stationary form here where on the site h you want to generate the values here. So, you will use this particular model and whence generate for site h, this has to be  $t + 1$  if you are using time period  $t + 1$ , and you are using the earlier value  $t$  and this is the cross correlation between the site j and site h. Just 1 second, I will do the correction correctly.

Let me explain this correctly again you are talking about the flow in site h in time period  $t$ , you are using the flows in time period in the station j in time period  $t$ . So, on the key site which is the site j here, sequence of observation is generated for site j using single site generation technique. So, what we have done? We have identified the key site which is the site j, use the single site model and generate the data you use either the stationary model or the non-stationary model. So, at j you have already generated the data. Now you move to the subordinate site. The subordinate site I am denoting it as h, now that is what we are doing now. So, this is your key site and this is what we denoted I am sorry it is not coming here.

(Refer Slide Time: 48:14)



Let me write it on a different sheet altogether. So, that this point is not missed. We have two sites now for the watershed and we want to generate data on both the sites. Let say this is the watershed, and this is site number. We will call this as j, this is as h, and this is what we call it as key site and this is called as the subordinate site. Now the key site we choose based on the length of the data there is a longer length of data that is available on

the key site and therefore, we call it as key site and also the quality of the data is much better than that available at h.

So, for the key site we simply use the single site model, we use single site model and generate the data on this, and then when we come to h, the subordinate site we use the generated data on the key site which is a single site which is done using the single site model, we use that and then build the cross correlation and develop the model for this. So, this is what we are doing now.

So, we are talking about generation on the site h, which is a which is the subordinate site because we are talking about only two sites now for time period t, and this is the mean of the flows at site h, this is  $r_{j, h}$  which is the cross correlation at lag zero between the flows at site j and the flows at site h. This is cross correlation at lag j, this is the standard deviation of flows at h, standard deviation at flow flows standard deviations of flows at j and so on. So, we are using the flows at j, to generate the flows at h as all other things essentially remain the same.

Except now, the random terms that we are building here this is the random term,  $u_t$  is a random term, and these are known now  $s_h$  is the standard deviation and so on, when we put it in this form, this will preserve the mean, standard deviation, and lag one correlations, and also the cross correlation not the lag one correlations it will also preserve the cross correlations. Let us see what are the properties of  $u_t$ , now this I have indicated here for use on the key site j; so, on the key site j you just use it except that the notations are different here i is the year and j is the month within this year. So, this is the single site model, this single site model we use for the site j, and then using the observations on site j you generate on site h.




(Refer Slide Time: 51:08)

### Multivariate Stochastic models

where  
 $u_t$  is a standardized random variate adjusted to incorporate the serial correlation at site h.

$$u_t = \zeta \frac{(X_{h,t-1} - \bar{x}_h)}{s_h} + t_t \sqrt{1 - \zeta^2}$$

$t_t$  is a standardized random variate

$$\zeta = \frac{r_h(1) - r_j(1)r_{j,h}^2(0)}{\sqrt{1 - r_{j,h}^2(0)}}$$


Now the  $u_t$ , which is a standardized random variate, we do this to incorporate the serial correlation at site h, serial correlation or the auto correlation. So, at site h you also want to build in the auto correlations lag one auto correlation. So, the  $u_t$  is defined as  $\zeta \times (X_{h,t-1} - \bar{x}_h) / s_h + t_t \sqrt{1 - \zeta^2}$ . So, at the site h you are doing it minus  $\bar{x}_h$  this is the mean of that, divided by the standard deviation plus  $t_t$  is the standard normal standardized random variate, remember we are still not talking about normal variant, normal variates. It is just a standardized random variate, and root of 1 minus  $\zeta^2$ , now this is developed by Fiering in 1969, and this has been discussed in the text book by Haan.

So, we will not go into the details of these we will just see how to use these expressions and the  $\zeta$  is given as  $r_h(1) - r_j(1)r_{j,h}^2(0) / \sqrt{1 - r_{j,h}^2(0)}$ . So, when I have only one subscript here that indicates the serial correlation. Serial correlation at lag one at site h, similarly the serial correlation of at lag one at site j, and this is  $r_{j,h}(0)$  which is the cross correlation between the flows at site j, and the flows at site h at lag zero, that is a cross correlation and then we are taking the square of that  $r_{j,h}^2(0)$  divided by root of 1 minus  $r_{j,h}^2(0)$ . So, this is how you compute  $\zeta$ . This is how **this is how** you can compute these and therefore, you compute  $\zeta$ , once you get  $\zeta$  you compute  $u_t$ , where you use  $t_t$  as a standardized normal variate, standardized random variate I am sorry not normal.

Now with this now, we form the multisite Markov model. Just look at this now we complete up to certain point now. So, starting with the first order Markov model which are the single site model, we have now develop a two site model for the two sites what we do? we identify one of the sites as a key site, and generate the data using the single site model for the key site. The other site which is the subordinate site, we build the model by taking into account the cross correlations among the key site and the subordinate site.

Where did we build the cross correlations? this is the term which has cross correlation  $r_{j,h}(0)$ , it also appears here, but also in the term  $u_t$  we have the lag one serial correlations. So, we are talking about both the cross correlations among these two sites as well as the serial correlation for the site for which you are generating it, and this is how the dependence on the other site has been brought into the Markov model, and as I mentioned in the term  $u_t$  here you have the lag one auto correlations of both the site  $h$  as well as site  $j$  and you also have the cross correlations between the site  $j$  and  $h$ .

This is how you build models for two sites I mean what I mean by that is when there is a cross correlation between the two sites, you will use these two site model to generate the data. Now this can be extended to multiple sites we will see how it is done in the next lecture. So, in today's lecture essentially we have started with the single site models and then went on to develop models for multiple sites, and the point that you must remember is that when typically you know these cases arise when there are several streams.

Let say  $X_1, X_2, X_3$ , etcetera these are the flows on different streams, and you want to generate data simultaneously on the random variable  $X_1, X_2, X_3$ , and so on. So, when you want to do this, if this can be assume to be independent then the single site models can be used we have discussed the single site earlier single site models earlier those can be used, typically the Thomas Fiering model in its stationary form or non-stationary form can be used.

However, when there is a dependence or the cross correlation exist significant cross correlation exist between 2 or more of these random variables, then we go for multivariate stochastic models, where the cross correlations among several of these can be brought into the models. We have also seen the cross correlation at lag  $k$  between two stations or between the random variables at two stations, let say  $j$  and  $k$  I am sorry  $j$  and  $h$  there are

two stations and then we are talking about cross correlations at a particular lag  $k$ , this is similar to what we do in the auto correlations by lagging time except that your lagging time on the other station now, and we have seen in the today's class how to compute the cross correlations at lag one through an example.

The same procedure can be adopted to obtain the cross correlations at lag  $k$ , and these cross correlations at lag, at different lags can be brought into our Thomas Fiering type of models, and specifically we have discuss today for two sites, how do we generate data by incorporating the cross correlations? initially we identify one of them as the key station simply use the single site model for the key station, and then go to the second station which is the subordinate station, and then build the cross correlations between this key station and the subordinate station the variables have these two places, and then build the Markov model. We will take this discussion further, and then look at how we generalize this for multivariable generation? So, thank you for your attention we will continue the discussion in the next class.