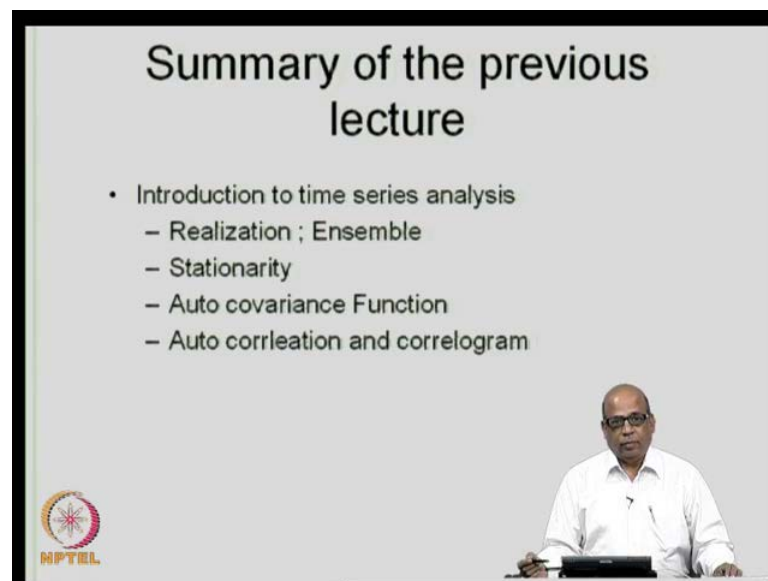


**Stochastic Hydrology**  
**Prof. P. P. Majumder**  
**Department of Civil Engineering**  
**Indian Institute of Science, Bangalore**

**Lecture No. # 11**  
**Time Series Analysis – II**

(Refer Slide Time: 00:23)



Good morning and welcome to this the lecture number eleven of stochastic hydrology. In the last class, we introduced the time series analysis, especially we talked about what is a realization, and what is an ensemble, and what are time properties, and ensemble properties, and so on. Then we introduced the important concept of stationarity of time series. If you recall the time series is suppose to be stationary, if the properties across the time remain the same. For example, the pdf of  $X_t$  is same as the pdf of  $X_{t+\tau}$  for all  $t$ , then we say that time series is stationary. We also introduced the concept of weak stationarity, we recon a time series to be weakly stationary of order  $f$ , if all the moments up to order  $f$  are the same across  $X$  across time  $t$  and  $t$  plus  $\tau$  for all  $t$ .

For example weak stationarity of order one indicates that the mean, which is the first moment is the same at time  $t$  and  $t$  plus  $\tau$  for all  $t$ . Similarity weak stationarity of order two indicates that both the mean and the covariance are the same at time  $t$  and  $t$  plus  $\tau$

for all  $t$ . Then we introduced the auto covariance function and the auto correlation and correlogram. Correlogram is also called as auto correlation function. We plot  $\rho_k$  which is of auto correlation at lag  $k$  with  $k$  on the X axis that plot is called as correlogram, and correlogram if you recall indicates the memory of the process, that is how far into the past does the process remember  $\rho_k$  indicates the dependence of  $X_t$  on  $X_{t-k}$  or  $X_t$  on  $X_{t+k}$  depending on how we compute it.

(Refer Slide Time: 03:02)



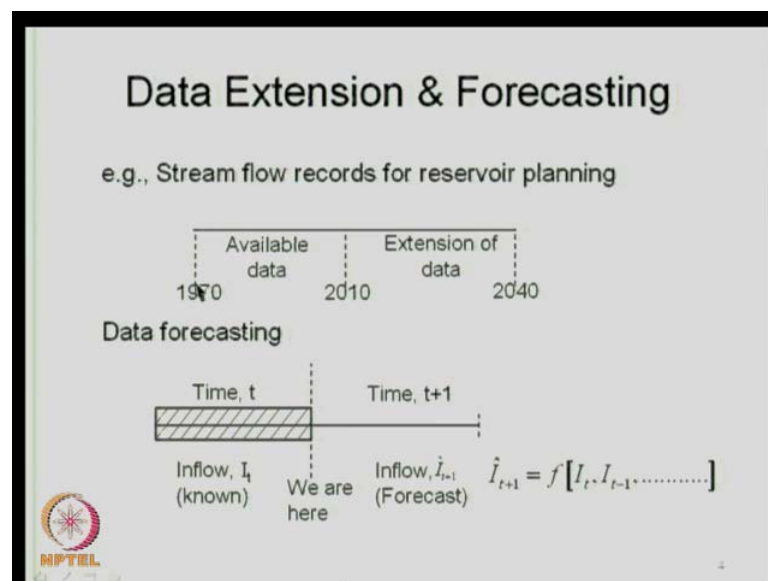
We today discuss an important an important aspect of stochastic hydrology, and we will introduce the concepts of data extension and forecasting. As I mentioned in the last class towards the end in any hydrologic planning and design exercise, we rely on the observed data. So, the observations are made over a period of time say for example you have 30 to 40 years of stream flow data at a particular location.

Using these observed data we want to make decisions on the designs that are going to serve us for the next so many years next 50 years, 100 years and so on. So the basis on which we make the hydrologic designs is the observed data sequence that we have. So, we have levels of two different types of problems here, where in the first case we will be talking about decisions that have long term implications; for example, you want to build a reservoir build a dam and you want to fix the capacity of the reservoir or how high should be the dam this decision. You will make base on the historical data, let say 50 years of past data monthly stream flow data and evaporation data or rainfall data etcetera,

you have observed data for last so many years and based on that you would fix the height of the dam this decision that you make based on the historical data has long term implication, because the dam will serve for next 100 years also. So, this is a decision that you are making now which has a long term implication on the other hand the problem of forecasting will have relatively short term implication, for example, you may have the data for last 5 years and you are standing at the beginning of June of this year and then you would like to forecast what you have likely have to be the monsoon flows in the reservoir or what is likely to be the rainfall in catchment area during this particular season and so on.

So in the forecasting essentially we are talking about short time duration relatively short time durations whereas, in the extension and generation. We will be talking about long time duration now both of these rely on the observed data and as said the main basis for both data extension as well as data forecasting is that a history provides a valuable clue to the future. So, we use the historical data extract the information from the historical data and then make an assessment of how the future is likely to be. So in today lecture, we will introduce both procedures for both data extensions as well as data forecasting, and then in the subsequent lectures, we will take this discussion forward when data extensions as I have just mentioned.

(Refer Slide Time: 06:16)



We would be looking at the available data. Let say from 1970 to 2010 as an example we have available stream flows for the last so many years about 40 years and we want to extend this record to let sat 2040, because we would like to make the decisions based on data from 1970 to 2040. So, this is the problem of extension of data the data generation will include extension of data procedures for extension of data, but additionally the data generation will provide us with several sequence is of the data, which has the same properties same statistical properties of the observed data by available data here mean that these are the data that are observed actually observed on the field data forecasting will be dealing.

Essentially with shot time spans, let say you are interested in forecast for the next 10 days or forecast for the next one month and so on. So, we will be using the data up to the time period  $p$ , let say we are here this may be June of a particular year. So we have the observed data with us until that particular time then we would like to forecast standing at the end of June of a particular month particular year. We would like to forecast what is like it to happen to the happen during the next time period, which is next month July. So this is the problem of data forecasting.

So data forecasting we provide as a function of the flows in the particular case we are talking about the inflow to a reservoir inflow which has been observed during the month  $t$  the previous month  $t - 1$   $t - 2$  and so on. So the entire history of this process up to this process is available with us using this entire history we would like to forecast what is likely to happen to flow during time period  $t + 1$  what is the expected flow during the next time  $t + 1$  that is a forecast. So we will today introduce methods of data forecasting as well as data extension.

(Refer Slide Time: 08:50)

### Data Extension & Forecasting

Calibration data      Test data

$X_1$        $X_{T-2}$   $X_{T-1}$   $X_T$        $X_{T+1}$        $X_N$


Use first 'T' values to build the model; rest of data to validate it

$F_{T+1}, F_{T+2}, \dots, F_N$ : forecasts obtained from the model

$(X_{T+1} - F_{T+1})$   
 $(X_{T+2} - F_{T+2})$   
 $(X_N - F_N)$

}

Forecast errors

5

Now, in the absence of any method, let say that we do not have any mathematical or statistical method available with us we all we have is the observed data and we would like to say this is the forecast for next time period. What are the procedures that we may use in to tell you we may simply say that just take the average of this and put it fore forecast for the next time period or simply look at the what has happened during the last period and say that the same thing you likely to continue for the next time period. So, like this intuitively we may do some forecast based on the observed data we all start with those kind of methods, but more formally what we generally do is that any particular method that we want to use or any particular model that we would like to build for forecasting we build it on first part of the data we call that as calibration data.

So, we build the model using the first part of this data this can be the 50 percent 70 percent or 80 percent of the available data and then we calibrate the model and test the model. So, calibrated on the remaining part of the data, let say you build a forecasting model using this data you calibrated that model and use that model for the next part of the data, because you already have this data you can then calculate the errors that you get from you get by using the model, what mean by that is that let say you have used up to time t you have used the data for calibrating the model use the model. So calibrated, first to forecast the flow during time period t plus 1, if you are talking about flow foresting model; so we get the forecast we denote it as F t plus 1, but we also have the observed

data for time period  $t + 1$ . So, the error will be  $X_{T+1} - F_{T+1}$ , it should have been actually  $X_{T+1}$ , but we have forecasted  $F_{T+1}$ .

Similarly, at the end of time period  $t + 1$ , we use actual value of  $X_{t+1}$  not the forecasted value use the actual value of  $X_{t+1}$  and then forecast it for  $X_{t+2}$ . So,  $X_{t+2}$  is actual value that has been realized, but  $F_{t+2}$  could have been your forecast. So, this gives you the errors in the forecast. So, we compute the errors as  $x_{t+1} - F_{t+1}$  etc. So, you can now form a series of the errors same type now these are some of the basics of what we do in the forecasting problem as just I mentioned. If we do not have any way of forecasting any formal algorithm by which we can forecast what we would have intuitively done is to simply take the averages. So, the first method of forecasting is simply based on the averages.

(Refer Slide Time: 12:08)

**Data Extension & Forecasting**

Method of simple averages: take the average of all the data up to period 'T' as the forecast for period (T+1)

$$\hat{X}_{t+1} = F_{T+1} = \frac{\sum_{t=1}^T X_t}{T}$$

$$\hat{X}_{t+2} = F_{T+2} = \frac{\sum_{t=1}^{T+1} X_t}{T+1} \text{ and so on}$$

For series with jumps, trends and periodicities this is not a good procedure

So, let us look at what we do this case let say we have the data up to time period  $t$ , if we are talking about monthly flows and we have the data for last about 50 years which means 600 values we have and we want to forecast what is likely what is the likely flow for the next time period that is 600 and first time period then the easiest way of doing it is take the long term mean that is mean of all the values that you have observed and put that mean to be the flow mean flow the flow that is likely to be happen during the next time period  $t + 1$ .

We proceed to next time period. So,  $t + 1$ , you would have got the actual value that have occurred take the values up to  $t + 1$  and forecast for  $t + 2$ . So, use all the values that you have at your disposal take the average of all those value and put that average as the forecast for the next time period this is the first level of forecasting. So, this is the method based on simple averages that is what is written more formally here whenever use a cap here it means that it is a forecast. So,  $X_{t+1}^c$ , which is the forecast for the time period  $t + 1$  also denoted as  $F_{t+1}$  this is simply equal to the sum of all the observed values divided by the number of time periods. So,  $t$  also indicates the number of time period in this case apart from indicating the specific time period here so, you take all the values up to that particular time period. Take the average of that and call that as a forecasted of so as you progress in time you will have more and more values to compute the average and that average you are putting it for as you are forecast.

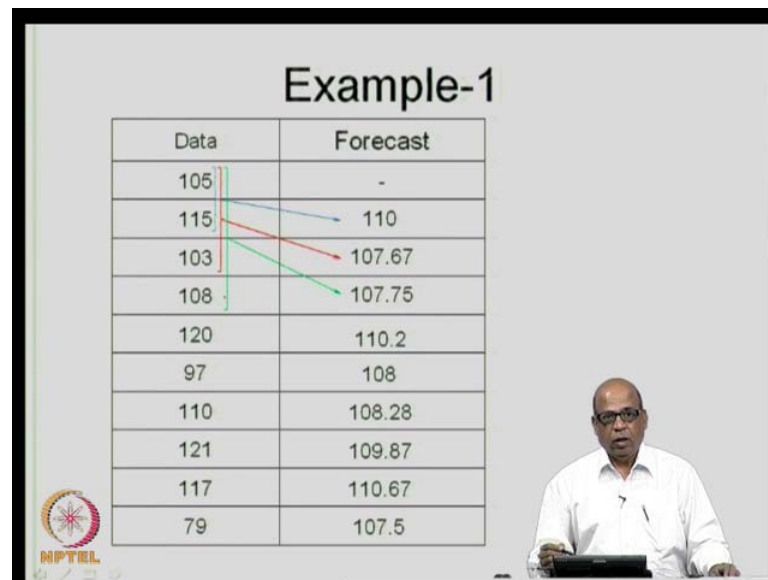
As you can see this method will not be useful where you have a strong periodicity or a trend etc, on the data if you have a smooth process. Let us say that you draw a smooth process let say your time series is something like this then your forecasting let say at this point based on the values we have got may be your forecast will be somewhere here and next time forecast will be somewhere here like this it may keep on going. So, if you have a smooth process this kind of average based procedure this kind of average based procedure will work reasonably all right whereas, if you have a time series let me show it here if you have a time series, which has performed like this and then suddenly there is either a jump or drop then your methods will fail for example, you may keep on getting forecast something like this then the forecast will be slowly rising like this.

So, you may not get a good forecast when you have either a jump or a periodicity or even a trend. So, if you have a trend in the time series something like this then your forecast using the average method will not work well a improvement of the forecast using the average method is not to take all the values, but to take the last few values which have occurred and then use that average is computed using the last few values and compute your forecast this we call as the moving average method; that means, we do not take all the values, but we keep a window of time fixed and take always the values within that window and then compute our averages.

(Refer Slide Time: 16:44)

### Example-1

Data	Forecast
105	-
115	110
103	107.67
108	107.75
120	110.2
97	108
110	108.28
121	109.87
117	110.67
79	107.5



So, first I demonstrate what mean by taking all the values and then computing the average and putting that as the forecast, let us look at these values this example we have the data these are the observed data 105, 115, 103 etcetera, like this observe here that here is a sudden drop in the data 120 that it was operating around certain level and suddenly there is a big drop here and there is a big jump again. So, this is the data how do we use this data for forecasting we have the observed data 105 and we want to forecast for the next time period, so from these 2 data 105 and 115. You forecast 110 then 105 , 110 and 103 you forecast 107.67 ; that means, you take the average of all of these and then forecast 107. 67 and. So, this forecast is actually for the next time period.

Then you take the all these averages and forecast 107.75. So, this has been got by taking the average of 105, 115, 103, 108 all of this values are used to forecast this forecast is for the next time period and so on. So, like this you get and keep getting the forecast for the next time period. So, the forecast that are shown here are for actually the time period subsequent to that for example, this is the forecast for this time period 103 and this is this forecast for this time period and so on.



(Refer Slide Time: 18:30)

**Data Extension & Forecasting**

Method of Moving Averages (MA)

The diagram shows three overlapping windows of length  $T$ . The first window covers time periods  $T$  to  $T+1$ . The second window covers  $T+1$  to  $T+2$ . The third window covers  $T+2$  to  $T+3$ . Each window is represented by a horizontal line with arrows at both ends, and the length of each window is labeled as  $T$ .

- As a new observation becomes available, new average is computed by dropping the oldest observation, including the newest one.
- No. of data points used for computing the average remains the same
- Uses the latest 'T' periods of known data

NPTEL

Now, as said instead of using all the values before we used the values during a fixed time period and call that as the moving average method. So, we keep shifting the time windows by one time step and then take the average. So, on the same time steps and call that as your forecast what mean by that is that, let say you have time period up to time period  $T$  use the time up to values up to time period  $t$  take the average and call it as forecast  $T$  plus 1.

In the next time period you would have got one more value added to that. So, discard the previous value and then take the average over the same time period  $T$ , let say  $t$  was 5 months and you are taking 5 months averages to forecast the next time period. So, when you move to the next time period you discard the oldest value and keep the latest 5 values to forecast the next value then you go to the next time step there are two time step that you will discard take the latest 5 months value average and then forecast for the next time period and so on. So, essentially what we are doing is the window of the time over which the average is taken is kept constant and this window is shifted every time by one time step and the average is taken over the previous  $T$  time periods and the average is taken as the forecast for the next time period. So, essentially we use the latest  $t$  time periods of known data and this  $T$  remains constant this window remains constant essentially, that means that the number of points data points that we use to compute the average remains the same across time as we progress on in the time.

(Refer Slide Time: 20:39)

### Example-2

Data	MA (3)	MA (3 x 3)
105	-	
115	-	
103	-	
108	107.67	
120	108.67	
97	110.33	108.89
110	108.33	109.11
121	109	109.22
117	109.33	109.33
79	116	

So, let us see what we mean by that. So, we are talking about moving average of the order 3 which means essentially we are taking the average for using 3 values 3 previous values. So, first 3 time period 105, 115 and 103, we use this take the average of that and then call that as forecast for the next time period 106.7. Similarly when we go to the next time period we discards this value and take the average of these 3 values 115, 103, and 108 and call it as the forecast for the next time period 108.67, next time period we discard these two values and take the next 3 values 103, 108, 120 and call that as forecast for this time period and so on. So, like this we get the forecast for all time period. So, essentially every time we are using 3 values 3 previous values 105, 115 and 103, which are the previous values to this and call it as for forecast for this time period. So, this time window remains the same every time we are taking previous three values and computing the average and calling it as the forecast

Now, we go one times one step ahead; that means, instead of reckoning the average only from the previous actual observed values we recon the average from the moving averages themselves; that means, now we will go one more order we take the averages of the moving averages themselves and then compute the forecast what mean by that we have the moving averages of order three. Now will use three of these moving averages and compute the average of the moving averages and call that as the forecast. So, this a smoothing process we are smoothing. So, like this we use these three moving averages and get the forecast for this why do write the forecast here. Because these 3

moving averages have used values up to this point up to 120 and this will be the forecast for what has been recorded as 97. So, we would have forecasted 108.89 using these 3 moving averages like this next time we use these 3 moving averages to compute a forecast for this and so on. So, we calculate like this and get the forecast. So, this is moving average of 3 by 3 second order moving average.

So, these are some of the methods based on averages one is simply take all the historical data take the average and average as computed thus would be taken as the forecast for the next time period next level we take the moving average of the first order where we fix the window always take the data for those time for those number of time periods take the average and then put it as the forecast for the next time period then we as a demonstrate. It here we go to two orders of moving averages we take the averages of the moving average themselves for example, we take 3 moving averages take the average and call it as a forecast and so on. So, these are methods based on averages of data forecasting as said these methods will work well when the process is fairly smooth we do not have a very prominent periodicity, prominent trends increasing trends or prominent germs in the data so on. So, when we have a reasonably smooth process that methods based on averages work reasonably well.

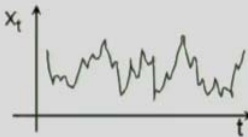
If you have data on monthly stream flows for example, these kind of methods may not work well, because monthly stream flows exhibits a significant periodicity there especially monsoon climates like ours where June month flow is serially correlated with may month flow or the June month of the previous year and soon. So, there may be significant periodicities in the data when you have significant periodicities or any significant deterministic component other than the mean then these methods will not work well.

(Refer Slide Time: 25:26)




### Data Generation – Uncorrelated Data

Purely random stochastic process:

Plot the time series



Plot the correlogram

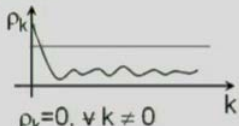


So, we will now look at data generation first we will consider uncorrelated data that is what? we discussed earlier as purely random stochastic process or purely stochastic processes if you see the correlogram for a purely stochastic processes the correlogram should indicate that data are uncorrelated if you have data which are uncorrelated where which will let say plot the correlogram rho k verses k and all the correlations that you have seen for k naught equal to 0 they are all insignificant they are all statistically insignificant, then the data is uncorrelated then what do we do.

(Refer Slide Time: 26:18)

### Data Generation – Uncorrelated Data

If the correlogram indicates that the time series is purely random




$\rho_k = 0, \forall k \neq 0$

Mainly used for flood peaks, storm intensities, short duration rainfall etc .

Not useful for stream flows, seasonal rainfall, and such long time processes.

- $X_t, X_{t-k}$  are independent
- Distribution of  $X_t$  is known
- Generate  $X_t$  using data generation technique to follow given distribution with parameters estimated from sample



12

If the correlogram indicates that  $\rho_k$  is equal to 0 for all  $k$  not equal to 0 what have shown here is the sample correlations sample auto correlations. So, there will not be exactly equal to 0, but there will be statistically insignificant. So, if you notice from the correlogram that for all  $k$  not equal to 0 the  $\rho_k$  are all insignificant statistically insignificant then you can say that the data is uncorrelated in which case if you are distribution for  $X_t$  is known, let say  $X_t$  follows a normal distribution or a long normal distribution or gamma distribution etcetera; so, if your distribution for  $X_t$  is known or can be estimated from the observed data of  $X_t$  along with this parameter then what? We do we have seen the methods data generation earlier using the specified distribution. If you recall what we do there, we set  $r_u$  which is the uniformly distributed random number in the interval 0 and 1 to be equal to  $F$  of  $X$  where  $F$  of  $X$  the if can write it here the  $F$  of  $x$  or let say if you recall we set  $F$  of  $Y$  is equal to  $R_u$  and then solve for  $Y$ .

So, this is what? we do there solve for  $Y$ .  $F$  of  $Y$  as indicated here is the distribution of the particular is the c d f of that particular distribution for example, exponential distribution if you have ten you compute  $F$  of  $Y$  and set it equal  $R_u$  where  $R_u$  is the random number uniformly distributed random number in the interval 0 and 1 and then solve for  $Y$  and in cases where you cannot explicitly solve for  $Y$  there are methods available for example, in the normal distribution we take the standard normal distribution and then you go to the tables, which provide you standard normal deviants and using that you estimate the  $Y$  you generate the values of  $Y$ .

So, this is what we do when you are sure that the data you have observed are all correlated that mean all uncorrelated; that means,  $X_t$  and  $X_{t-1}$  or  $X_{t-k}$ .  $X_t$  and  $X_{t-k}$  can be assumed to be independent. Then you use the distribution for  $X_t$  generate  $x_t$  distribution for  $X_{t+k}$  generate  $X_{t+k}$  and so on. So, each time period you generate different sets of values knowing the distributions of those particulars time of the variable of that particular time period, but in general in hydrology most of the data are serially correlated as just mentioned if you are looking at the stream flow during a particular month let say stream flow during August.

(Refer Slide Time: 29:35)

**Data Generation – Serially Correlated Data**

- Most hydrologic time series exhibit serial dependence e.g.,  $X(t)$  correlated with  $X(t-\tau)$

$$\rho_k \equiv (\rho_1)^k$$
$$\rho_k \rightarrow 0, k \rightarrow \infty$$

Exponentially decaying

First order Markov

MPTEL

The slide features a presenter in a white shirt and glasses standing behind a podium. The background is a light gray with a black border. The text and equations are in black. The graph shows a curve starting at a high point on the y-axis and decaying towards the x-axis. The y-axis is labeled  $\rho_k$  and the x-axis is labeled  $k$ . The text 'Exponentially decaying' is written next to the curve. The text 'First order Markov' is written below the graph. The MPTEL logo is in the bottom left corner.

This is serially correlated with what has happened during the month of July not only what has happened during the month of July, but also what has happened in June month also what has happened during August month of the previous year what has happened during August month and July month of the previous year and so on. So, the data will be serially correlated the value that have you have realized during particular month will be dependent on what has been the value of previous month what has been the value of the same month during the previous year during the two years before that and so on. So, these are called as serially correlated.

So, we will see first a method based on lag one correlation when we have significant dependence on what has happened during the previous month and we introduce a concept of first order Markov process. Where, if you have a  $\rho_k$  which is exponentially decaying like this; that means,  $\rho_k$  can be approximated as  $\rho_1$  to the power  $k$   $\rho_1$  is the lag 1 correlation to the power  $k$  and  $\rho_k$  tends to 0 as  $k$  tends to infinity. So, for theoretical first order Markov process you will get a correlogram, which has a exponentially decaying shape like this. So, because of most hydrologic time series are serially correlated that is  $X(t)$  is correlated with  $X(t-\tau)$  we use several methods which are based on a data generation which are based on serially correlated data and the first one that which we are introducing very commonly used in hydrology is that based on first order Markov processer.

When you have this, what does it mean? This means that the first lag 1 correlation; lag 1 correlation is much more significant compared to any other lag correlations. So, as you can see  $\rho_2$  will be much smaller than  $\rho_1$   $\rho_3$  will be much smaller than  $\rho_1$  and so on. So in fact, any correlation at any lag can be obtained from the correlation of lag 1, itself following this notation this also indicates that the memory of the process is short. So, you may have one time step memory two time step memory and so on.

That means if you have  $X_t, X_{t+1}, X_{t+2}$  and so on,  $X_{t+k}$  depends on  $X_{t+k-1}$  or  $X_t$  depends on  $X_{t-1}$ . If you **if you** have the process, where memory is only to the limited extent of the previous time period, then you can use the first order Markov process more formally it is defined as will anyway introduce the formal definition of Markov chain later on, but more formally we denote a first order Markov chain as if you have a probability of  $X_t$  given  $X_{t-1}, X_{t-2}$  and so on up to  $X_{t-naught}$ , which is the complete history of the process if we can approximate as probability of  $X_t$  given  $X_{t-1}$  then we use the first order Markov chain or Markov process in this particular process. What does this mean? This means that the entire information of the history of the process given by  $X_{t-1}, X_{t-2}$  etc, up to  $X_{t-naught}$  is all contained in the previous example  $X_{t-naught}$  itself.

So, the conditional probability of  $X_t$  given  $X_{t-1}, X_{t-2}$  etc,  $X_{t-naught}$  can be approximated by probability of  $X_t$  given  $X_{t-1}$  then this we call as the first order Markov chain. In general in hydrology we can approximate especially the large time processes for example, monthly stream flow seasonal stream flow etcetera, by the first order Markov process, and we introduce the data generation models for the first order Markov processes as indicated here. So, essentially what we are looking at let say we are looking at the annual stream flow as said.

(Refer Slide Time: 34:48)

The slide is titled "Data Generation – Serially Correlated Data". It describes a first-order Markov process with the equation  $X_{t+1} = \mu_X + \rho_1 (X_t - \mu_X) + \varepsilon_{t+1}$ . The term  $\mu_X + \rho_1 (X_t - \mu_X)$  is labeled as the "Deterministic component", and  $\varepsilon_{t+1}$  is labeled as the "Random component". Below the equation, it states  $\varepsilon \sim \text{Mean 0 and variance } \sigma_\varepsilon^2$ . At the bottom, it says "This model is stationary w.r.t both mean and variance". A speaker is visible in the bottom right corner of the slide frame, and the NPTEL logo is in the bottom left.

If you look at large time steps monthly time steps annual time steps and. So, on and you are able to disregard or negate periodicities inherent in the data; that means, if the periodicities are not very strong and you are you can afford the periodicity then we introduce the stationary Markov process as follows  $X_{t+1}$  is equal to  $\mu_X$  which is the mean of the process plus  $\rho_1$ , which is the correlation of  $X_t$  and  $X_{t+1}$  which indicates the dependence of  $X_{t+1}$  on  $X_t$  into  $X_t - \mu_X$ .  $\mu_X$  is the stationary mean plus a random component that is all very simple generating process. So, this  $X_{t+1}$  the value you are generating using the data that is available with you will depend on  $X_t$  through  $\rho_1$ . So, the dependence of  $X_{t+1}$  on  $X_t$  is measured by  $\rho_1$  and  $\mu_X$  is the long terms mean of the process, which remains constant. So, we are taking the process to be stationary in mean. So, when we use this and we introduce the random component with the mean zero and its own variance  $\sigma_\varepsilon^2$ .

So, this random component has a mean of zero and its own variance  $\sigma_\varepsilon^2$  and we would like to generate using this model generate those values. Which will have the same mean as  $\mu_X$  and the same variance as  $\sigma_X^2$ , because we would like to have the same mean the epsilon must have the mean of zero as we after representing demonstrate, but to maintain the variance the same the  $\sigma_\varepsilon^2$  must have a particular variance we will just see what happens to that. So, this model essentially generate  $X_{t+1}$  that is the sequence  $X_t$  is to generates using the previous value of  $X_t$  and the dependence of the current value on the previous value given by  $\rho_1$ . Now as




said the properties of epsilon t plus1 in this case are important epsilon t plus1 must have a mean of 0 the variance of sigma e square.

(Refer Slide Time: 37:59)

**Data Generation – Serially Correlated Data**

$$\begin{aligned}
 E[X_{t+1}] &= E[\mu_x + \rho_1 (X_t - \mu_x) + \varepsilon_{t+1}] \\
 &= E[\mu_x] + \rho_1 \{E[X_t] - E[\mu_x]\} + E[\varepsilon_{t+1}] \\
 &= \mu_x + \rho_1 (\mu_x - \mu_x) + 0 \\
 &= \mu_x
 \end{aligned}$$

$$\begin{aligned}
 \sigma_x^2 &= E[X^2] - (E[X])^2 \\
 &= E[(\mu_x + \rho_1 (X_t - \mu_x) + \varepsilon_{t+1})^2] - (E[\mu_x])^2
 \end{aligned}$$




So, for this X t plus 1 given by this to have the same mean as mu X what does it has to satisfy we will take the model expected value. So, expected value of X t plus1 and this is the model mu X plus rho1 X t minus mu X plus epsilon t plus. So, that is expected value of mu X rho1 into expected value of X t minus expected value of mu X plus expected value of epsilon t plus1 and as have mentioned expected value of epsilon t plus 1 is 0. This is what we have used epsilon has a mean of 0. So, this value becomes 0 and this becomes 0, because this is mu X minus mu X and this is mu X itself. So, expected value of X t plus1 becomes mu X. So, the model as defined by this will have the same expected value as the long term mean mu X which is stationary.

Now, we want to see what happens to the variance of X t plus 1. So, X t plus 1 as defined by this, let us look at the variance. So, the variance is simply expected value of X square minus expected value of X whole square. So, will take expected value of X t plus1 the whole square here minus expected value X t plus 1 the square minus expected value X t plus 1 the whole square if we simplify this is the simplification then and we have used the fact that expected value X t plus 1 is equal to mu X itself.

(Refer Slide Time: 39:40)

### Data Generation – Serially Correlated Data

$$\begin{aligned}
 \sigma_x^2 &= E\left[\mu_x^2 + \rho_1^2 (X_t - \mu_x)^2 + \varepsilon_{t+1}^2 + 2\mu_x \rho_1 (X_t - \mu_x) + \right. \\
 &\quad \left. + 2\varepsilon_{t+1} \rho_1 (X_t - \mu_x) + 2\mu_x \varepsilon_{t+1}\right] - (E[X_{t+1}])^2 \\
 &= E\left[\mu_x^2\right] + \rho_1^2 E\left[(X_t - \mu_x)^2\right] + E\left[\varepsilon_{t+1}^2\right] + 2\mu_x \rho_1 E\left[(X_t - \mu_x)\right] \\
 &\quad + 2\rho_1 E\left[\varepsilon_{t+1}\right] E\left[(X_t - \mu_x)\right] + 2\mu_x E\left[\varepsilon_{t+1}\right] - (E[X_{t+1}])^2 \\
 &= \mu_x^2 + \rho_1^2 E\left[(X_t - \mu_x)^2\right] + E\left[\varepsilon_{t+1}^2\right] + 0 + 0 + 0 - \mu_x^2 \\
 &= \rho_1^2 \sigma_x^2 + \sigma_\varepsilon^2 \\
 \sigma_\varepsilon^2 &= \rho_1^2 (1 - \sigma_x^2)
 \end{aligned}$$


So, when we simplify this you get  $\sigma_x^2$  is equal to  $\rho_1^2 \sigma_x^2 + \sigma_\varepsilon^2$ . All of this simplification is given here; that means,  $\sigma_\varepsilon^2$  can be written as  $\sigma_\varepsilon^2 = \rho_1^2 (1 - \sigma_x^2)$ . What does this indicate? This indicates that if you want to have the same  $\sigma_x^2$ , the idea here is you would like to generate the data using this model and this model will have the same mean  $\mu_x$  as the historical data and for this model or the generated values to have the same variance as the historical  $\sigma_x^2$ ; you mean or your standard deviation or the variance of the  $\varepsilon_{t+1}$  must be given by  $\rho_1^2 (1 - \sigma_x^2)$  that is the idea there. So, you must use your random components with 0 mean and variance given by this as you know as you can see  $\rho_1$  is given from the historical data for you can estimate the data  $\rho_1$  and  $\sigma_x^2$  is the stationary variance of the time series, which is also known.

(Refer Slide Time: 41:20)

The slide contains the following text and formulas:

**Data Generation – Serially Correlated Data**

If  $X \sim N(\mu_x, \sigma_x^2)$  then  $\varepsilon \sim N(0, \sigma_\varepsilon^2)$

If  $\{u_t\} \sim N(0, 1)$ ,  $\{u_t \sigma_x \sqrt{1 - \rho_1^2}\}$  is  $N(0, \sigma_\varepsilon^2)$

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

Standard normal deviate

First order stationary Markov model  
Or  
Thomas Fiering model (Stationary)

MPTEL

A presenter is visible in the bottom right corner of the slide.

Now, a specific case where if  $X$  follows normal distribution that is you are observed stream flow data for example, stream flow data follows normal distribution with mean  $\mu_x$  and  $\sigma_x^2$  then  $\varepsilon$  should be also normally distributed with zero mean and  $\sigma_\varepsilon^2$  as the variance. So, we now introduce if  $u_t$  has a normal distribution of 0 1 then  $u_t$  into  $\sigma_x \sqrt{1 - \rho_1^2}$  that is  $u_t$  into  $\sigma_x \sqrt{1 - \rho_1^2}$  is normally distributed with 0 mean and  $\sigma_\varepsilon^2$ . So, this is what we want we will introduce  $u_t$  into  $\sigma_x \sqrt{1 - \rho_1^2}$  or  $u_t$  into  $\sigma_x \sqrt{1 - \rho_1^2}$  into the model. So, that we are ensuring that the random component will have a 0 mean and a variance of  $\sigma_\varepsilon^2$ .

So, we write the model now as  $X_{t+1} = \mu_x + \rho_1 (X_t - \mu_x) + u_{t+1} \sigma_x \sqrt{1 - \rho_1^2}$  why do we write this we are writing to ensure that the sequence we generate will have a normal distribution with 0 mean that is this random term will have a 0 mean and a variance of  $\sigma_\varepsilon^2$  that is why we are introducing this thing what is this  $u_{t+1}$  it is the standard normal deviate or the standard normal number random number with it is a random number which follows the standard normal distribution which means 0 means an unit variance. The model we thus write is called as the first order stationary Markov model first order because we are generating  $X_{t+1}$  using the previous value  $X_t$  only and not  $X_{t-1}$   $X_{t-2}$  and so on. So, have just choosing one previous value. So, it is one order Markov model it is stationary, because we are using the same

value of  $\mu_X$  and  $\sigma_X$  for all the values. So, it is stationary both in mean as well as standard deviation.

The first order Markov model is very popular in the hydrologic data generation especially when we are talking about annual stream flows the Markov model is straight away use and annual stream flows as I have mentioned are very useful in making designs hydrologic designs, for examples fix the capacity of the reservoir you would require annual stream flows and you generate the data using this particular model and use those generated sequence. Generated sequence to make decisions on reservoir capacity and. So, on. Now how do e generated values using this from the historical observed data you would have estimated  $\mu_X$  by your  $\bar{X}$  or  $\bar{X}$  then you can also calculate the  $\rho_1$  which is the lag 1 correlation coefficient. So,  $\mu_X$  is known similarly  $\sigma_X$  is known and  $\rho_1$  is known.

We start the process by assuming a value of  $X_t$ , let say you want to generate  $X_{t+1}$  to start kick start the process you want to generate  $X_{t+1}$ . So, first you assume  $X_t$  remember. Once you calculate this parameter this moments  $\mu_X$   $\rho_1$  and  $\sigma_X$  you forget about the data you have to only deal with this moment  $\mu_X$  and  $\sigma_X$  and  $\rho_1$ . So, first you assume  $X_t$  and typically it is assumed to be equal to  $\mu_X$ , itself the mean which is known. So, that this term becomes 0 and calculate take the standard normal deviant from your tables or any spreadsheets that give this normal deviants straight away or you can use your calculators to get uniformly distributed random numbers and then convert them into normally distributed random numbers or you simply write a math lab very simple math lab codes math lab functions you use to generate standard normal deviants.

So, every time for example,  $X_{t+1}$  you are starting  $X_t$  you have assumed to be equal to  $\mu_X$ . So, this term becomes 0  $\mu_X$  is a long term mean and  $\sigma_X$  is known  $\rho_1$  is known and this standard normal deviant you pick it up from the table or any of functions that are available. So, you get  $X_{t+1}$ , from  $X_{t+1}$  you generate  $X_{t+2}$ . So, here becomes  $X_{t+1}$  the value. So, that is just now generated and every time you change  $X_t$  as well as the standard normal deviant all other terms remain the same you generate the next value and so on, because you started the process by assuming a certain value for  $X_t$ . When you generate this numbers for a long fairly long sequence for fairly long sequence of numbers are generated. Discard the first few values to do away with the effect of the

initial value that you assume and use the remaining number of values. Remember there are two major assumptions here one is that the process  $X_t$  follows normal distribution with mean  $\mu_X$  and standard deviation  $\sigma_X$  and that the process is stationary in mean as well as standard deviation; that means, as you change from  $X_t$  to  $X_{t-1}$ , it will have the same mean  $\mu_X$  and same standard deviation  $\sigma_X$ .

(Refer Slide Time: 47:45)

**Data Generation – Serially Correlated Data**



If  $X \sim N(\mu_x, \sigma_x^2)$  then  $\varepsilon \sim N(0, \sigma_\varepsilon^2)$

If  $\{u_t\} \sim N(0, 1)$ ,  $\{u_t \sigma_\varepsilon\}$  (i.e.,  $u_t \sigma_x \sqrt{1 - \rho_1^2}$ ) is  $N(0, \sigma_\varepsilon^2)$

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

Standard normal deviate

First order stationary Markov model  
Or  
Thomas Fiering model (Stationary)

Let us see as just mentioned we generate a large set of values, let say if you have the observed data of 30 to 40 years you generate sequences of 100 years one sequence of 100 years another sequence of 100 years and. So, on how do we generate different sequence is always by changing the random deviates here. So, by changing the random deviates you generate those many sequences. So, generate large number of sequences and every sequence you discard the first 50 to 100 values.

So, let say 50 years of sequence is monthly 50 years of sequence you generated you can discard the first 5 values first 10 values and so on or you have generated large sequence of 150 years, 1000 years, which have they are typically used for large scale simulations or water resource systems then you discard the first 50 years, 50 values standard values and so on; so that the effect of first initial value that you have used to start the process that effect dies down, because you are using the normal distribution it is likely and often it happens that you generate negative values, because your standard normal deviants you are picking up will typically go from minus 3 to plus 3 and therefore, your value that you

generate thus can be negative and if you are looking at hydrologic variables physical variables for example, stream flow rain fall and so on. These values cannot take negative values. So, essentially what we do is when we come across to a negative value generated by this model retain the negative value as such for generating the next value. So, when you are generating the sequence you retain the negative values as generated by this model, but when you are using them in the application let say the sequence have several negative values and this sequence you want to use it in the application for fixing the reservoir capacity and so on.


So, when you are using the sequence with negative values in the application just replace the negative values by 0. We will have ample opportunities to demonstrate the applications of this kind of models in actual decision making. Subsequently, but write now let us understand how to use this model to generate values. So, let us take jus about three to four values and see how we can generate.

(Refer Slide Time: 50:39)

**Example-3**

Consider the annual stream flow data (in cumecs) at a river for 29 years

S.No.	Data	S.No.	Data	S.No.	Data
1	1093.31	11	1042.33	21	1444.97
2	1636.87	12	1492.13	22	1203.08
3	1485.67	13	1205.90	23	910.73
4	1579.51	14	1245.77	24	883.59
5	1443.00	15	1197.81	25	970.98
6	1327.40	16	1754.55	26	1001.92
7	1108.70	17	1108.56	27	1434.91
8	928.10	18	957.64	28	1635.00
9	840.83	19	1425.80	29	1875.78
10	1447.03	20	1128.62		



So, look at this example 29 years of annual flow is available and that data is shown here. So, this is annual stream flow data at a river for 29 years is available, this is the year number. So, like this 29 years and this is the stream flow data. So, we want to generate values from this stream flow data, if these were uncorrelated what we would have done we would have simply fit a distribution, and then use the methods for uncorrelated data as we have discussed in the earlier classes and generated values from this. But because

this is correlated that is that is the serial correlation associated with this particular data. We use the data that we have just described we will use the first order stationary Markov process and then generate the data to do that what are the first steps first is to estimate the mean standard deviation and lag 1 correlation for this. So, we will estimate that.

(Refer Slide Time: 51:40)

**Example-3 (contd.)**

For the data,  
 $\mu_x = 1269, \sigma_x = 281, \rho_1 = 0.255$

$$c_1 = \frac{\sum_{t=1}^{n-1} (x_t - \bar{x})(x_{t+1} - \bar{x})}{n}$$

$$r_1 = \frac{c_1}{c_0}; \quad c_0 = \sigma_x^2$$

Assume  $X_1 = \mu_x = 1269.33$

$$X_2 = \mu_x + \rho_1(X_1 - \mu_x) + u_{t+1}\sigma_x\sqrt{1-\rho_1^2}$$

$$= 1269 + 0.255(1269 - 1269) + (-0.464)281\sqrt{1-0.255^2}$$

$$= 1143$$

$$X_3 = 1269 + 0.255(1143 - 1269) + (0.335)281\sqrt{1-0.255^2}$$

$$= 1328$$

So, mean is estimated as 1269 sigma X 289 and the lag 1 correlation is got by the covariance of X t with X t plus 1 and the variance itself. So, lag 1 correlation is estimated by r 1 is equal to C 1 by C naught, because that rho k is gamma k by gamma naught where gamma k is a covariance between X t and X t plus k and gamma naught is the variance itself. So, using this we estimate row 1 actually the sample estimates are indicated by r 1. So, rho 1 in this case is a sample estimates, which here come out to be .255. So, to use the Markov model all you need these three parameters that is mu standard deviation and the lag 1 correlation. Once you estimate this you completely forget about the data all the information contained that is necessary for you to use the first order stationary Markov process is all available here.

So, start the generation process we first assume X 1 the first value equal to be mean which is actually 1269.33 here. It is written as 1269, but actual value is 1269.33 or we will use this as 1260 itself. So, we will use this as 1260 and. So, X 2 is generated as mu X plus rho 1 into X 1 minus mu X plus u t plus 1, which is standard normal deviant sigma X root of 1 minus rho 1 square. So, u t plus 1 is we can write it as u 2 here. So, mu

$X_2 = 1269 + \rho_1 (X_1 - 1269) + \sigma_1 \epsilon_1$  where  $\rho_1 = 0.255$  and  $\sigma_1 = 281 \sqrt{1 - \rho_1^2} = 281 \sqrt{1 - 0.255^2} = 268.464$ . So,  $X_2 = 1269 + 0.255(1269 - 1269) + 268.464 \epsilon_1$ . So, once you get  $X_2$  you use  $X_2$  to generate  $X_3$ . So, where does  $X_2$  come  $X_2$  comes in this place. So, this place you use  $X_2$  this is  $X_2$  again. So,  $X_3 = 1269 + 0.255(X_2 - 1269) + \sigma_1 \epsilon_2$  and every time you change this number. So, every time you changing this number as well as this number all other things remain the same all other term remains the same. So, you get  $X_3$  is equal to 1328 then using  $X_3$  you generate  $X_4$ . Now  $X_4$  this number has changed and this number has changed all other number remains the same. So, you will get  $X_4$ .

(Refer Slide Time: 54:32)

**Example-3 (contd.)**

$$X_4 = 1269 + 0.255(1328 - 1269) + (-0.051)281\sqrt{1 - 0.255^2}$$

$$= 1270$$

$$X_5 = 1269 + 0.255(1270 - 1269) + (1.226)281\sqrt{1 - 0.255^2}$$

$$= 1602$$

So, similarly  $X_5$  you use  $X_4$  to generate that. So,  $X_4$  was here. So, you use this number changes this number changes and you get  $X_5$  and so on. So, like this you can generate 1000 of values using the first order Markov process. So, essentially what we did in this lecture is will complete that. So, you generate such number of large values let say you wanted to generate 1000 years value 1000 year flows why do we require such large sequence is we need such larger sequence to simulate the performance of the system it does not mean that the system itself full exist for the 1000 years. We want to see how the system behaves if you have a sequence of 1000 years of flows how the system behaves. So, system performance we would like to estimate. So, we typically require large sequences of data and the first method that we have introduced to generate large sequences of data when the data are serially correlated is a first order Markov process



and it is famously called in hydrologic literature as the Thomas firing model after the people who have proposed this model.

Thomas and firing famous hydrologist this was proposed in the cities 19 cities. So, typically in hydrologic data generation Thomas firing model or the first order stationary model is the most commonly used data generation method the limitations of this model is that it assumes normal distribution that is the first one and it also assumes stationarity in mean and standard deviation. The values that you, so generate using the first order Markov process should preserve the mean of the data standard deviation of the data and also the lag one correlation of the data. We will see through certain applications large case studies etcetera how this is done. So, essentially in this lecture we have introduced the data forecasting methods and the data and one of the data generation methods.

In the data forecasting methods we have introduced the methods based on the averages. So, take the average of all the data available up to this point of time and then call it as forecast that is the first simple method and then the moving average method where you keep the window over which you take the average that window you fix and then keep moving the window itself. So, that is called as moving average number and in the moving average method you have higher order of moving averages; that means, you take first moving averages and then take the averages of the moving averages themselves to take to get the second moving averages and so on. Now, this methods as you as mentioned in the class will work well for fairly smooth processes that those processes which do not have a sudden jump in the data or do not have a significant periodicities significant trend increasing trend or decreasing trend etcetera when you have processes like that these methods will work well.

Then we introduce the method for data generation or the first order Markov process and in the specifically we have talked about stationary Markov process in the next lecture we will continue this discussion by first introducing non stationarity in the first order Markov model.