Stochastic Hydrology

Prof. P. P. Mujumdar

Department of Civil Engineering

Indian Institute of Science, Bangalore

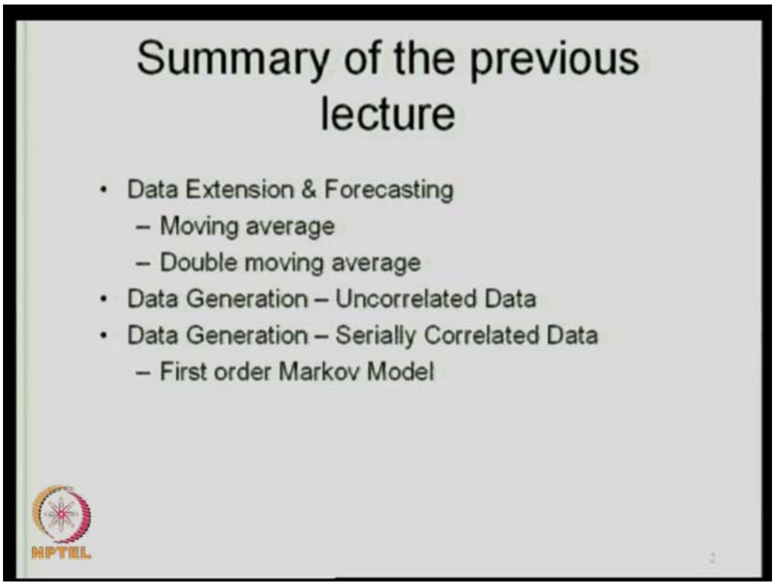
Module No. # 04

Lecture No. # 12

Time Series Analysis – III

Good morning and welcome to this the twelfth lecture of the course stochastic hydrology.

(Refer Slide Time: 00:25)



Summary of the previous lecture

- Data Extension & Forecasting
  - Moving average
  - Double moving average
- Data Generation – Uncorrelated Data
- Data Generation – Serially Correlated Data
  - First order Markov Model

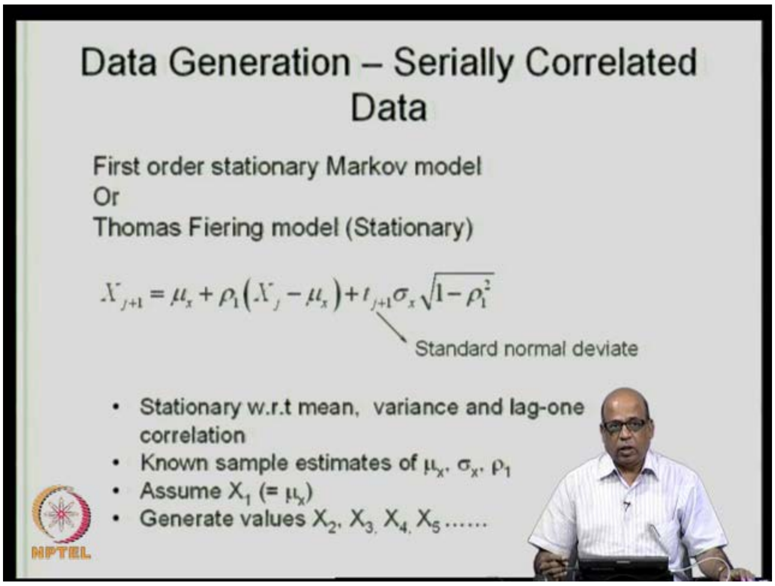
If you recall in the last lecture, we discussed methods about methods of data generation and forecasting and for the forecasting, we discussed specifically the average based methods that is the methods based on averages. Specifically, the moving average and the double moving average, where the time window is kept constant, and then you keep on computing the averages for over that time window and the window itself keeps on shifting across the data. And in the double moving average, what do we do? We take the moving average of the first order, and then take the averages of the moving averages themselves in a fixed time window.

And we discussed, the example of moving average of order three and moving average of order three by three into three; that is the second order moving averages; again also of order three. Then we discussed the data generation methods; in one of the earlier lectures, we had discussed data generation of uncorrelated data, where we used the distributions of the particular data, and then generate data from that particular distribution, by using the cdf and equating the cdf value to a uniformly distributed random number.

Then, in the last lecture, we also discussed data generation methods for serially correlated data, as I mentioned in the last lecture, we have in hydrologic data many times the data are serially correlated. For example, the July month's flow may be serially correlated with June month flow, and also correlated with the July month flow of the previous year. So, to account for such serial dependents in the data, we have discussed in the last lecture in the first order Markov model and specifically for generating annual stream flows the first order Markov model is also in hydrologic literature, often time called as Thomas fiering model, after the hydrologist who proposed this particular model.

So, we will continue the discussion today on serially correlated data, remember the assumption that we made in the first order Markov model. In the way, it was presented in the last lecture was that the flows are normally distributed and this is a stationary model; as we introduced in the last lecture this was a stationary model in the sense that the mean and the standard deviation and the lag one correlation all of which appear in the particular model they are all stationary across time.

(Refer Slide Time: 03:33)



Data Generation – Serially Correlated Data

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

$$X_{j+1} = \mu_x + \rho_1 \left(X_j - \mu_x \right) + t_{j+1} \sigma_x \sqrt{1 - \rho_1^2}$$

Standard normal deviate

- Stationary w.r.t mean, variance and lag-one correlation
- Known sample estimates of μ_x, σ_x, ρ_1
- Assume X_1 (= μ_x)
- Generate values X_2, X_3, X_4, X_5……

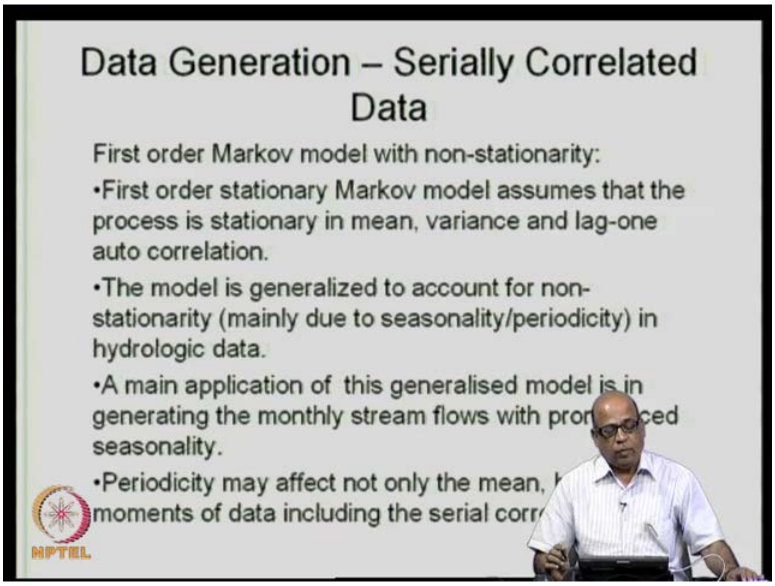
So, we will continue the discussion today as we wrote the first order model; first order Markov model it is X j plus 1 is equal to mu x, which is a stationary mean then the lag one correlation in to X j minus mu x plus that is a standard normal deviate t j plus 1 sigma x root of 1 minus rho 1 square, rho 1 is the lag one correlation, this is stationary with respect to mean variance and lag one correlation.

What we do in using this model, if you recall is from the data we estimate the parameters, we estimate the moments actually mu x and sigma x and also we estimate the lag one correlation rho 1. We use these sample estimates into this particular model, to start the model we assume a particular value of X j here. Let’s say X 1, we assume and then generate X 2. X 1 we assume and for convenience, we often assume X 1 to be mu x itself. So, that this term vanishes. So, starting with that we generate several values typically, we generate a large sequence of values from this particular model and discard the first few values. Let’s say, you generated for 200 years then first about 50 to 60 values you discard to make sure that the effect of the initial value assumption dies down.

And we generate sequences like X 2, X 3 etc after assuming the value for the first flow X 1. So, assuming the value of the first flow, we generate large number of such values. When you generate values using this model, the generated model, the generated data will have the same mean approximately the same mean as a historical mean, the same standard deviation as a historical standard deviation and the same lag one correlation approximately as a historical lag one correlation. So, once we generate the data you compare the historical mean standard deviation and lag one correlation with the generated mean standard deviation and lag one correlation.

These three comparisons must be acceptable, if there are certain cases where the model does not perform well either in terms of the standard deviation or in terms of lag one correlation, it then means that the assumptions that we have made in building this model namely, that it is a stationary model and that the flows, follow a normal distribution. These assumptions May not be valid for the data that you are using. So, a test for the model to be valid is that all the three namely the mean, the standard deviation and the lag one correlations must compare well, between this historical data and the generated data.

(Refer Slide Time: 07:08)



Data Generation – Serially Correlated Data

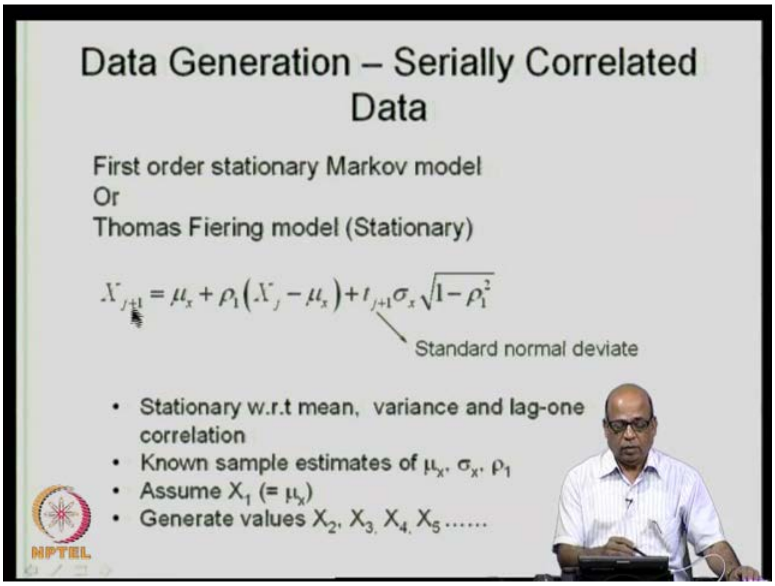
First order Markov model with non-stationarity:

- First order stationary Markov model assumes that the process is stationary in mean, variance and lag-one auto correlation.
- The model is generalized to account for non-stationarity (mainly due to seasonality/periodicity) in hydrologic data.
- A main application of this generalised model is in generating the monthly stream flows with pron ced seasonality.
- Periodicity may affect not only the mean, moments of data including the serial corr

Now, we will start relaxing the requirement of the stationarity a bit, what did we assume in the Markov model that we just presented that the mean remains the same; that means between June month, July month etc we are not changing the mean. We have a time series let’s say, for the last fifty years you have collected the flows and you have that particular time series, you have one mean for the same for the entire time series, you have one standard deviation and so on.

So, these are stationary mean, stationary time stationary standard deviations and so on. But as we are well aware the hydrologic time series exhibit non stationarity, especially when you are talking about the stream flows. Let’s say the stream flow of June month will have its own mean, which will be much different from say flows during April or flows during February and so on especially in the monsoon climate. So, there are many situations where the mean standard deviation and the lag one correlations will be significantly different from one month to another month and therefore, it is essential that we build in this variation or this non stationarity in the moments into the generating model.

(Refer Slide Time: 08:48)



Data Generation – Serially Correlated Data

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

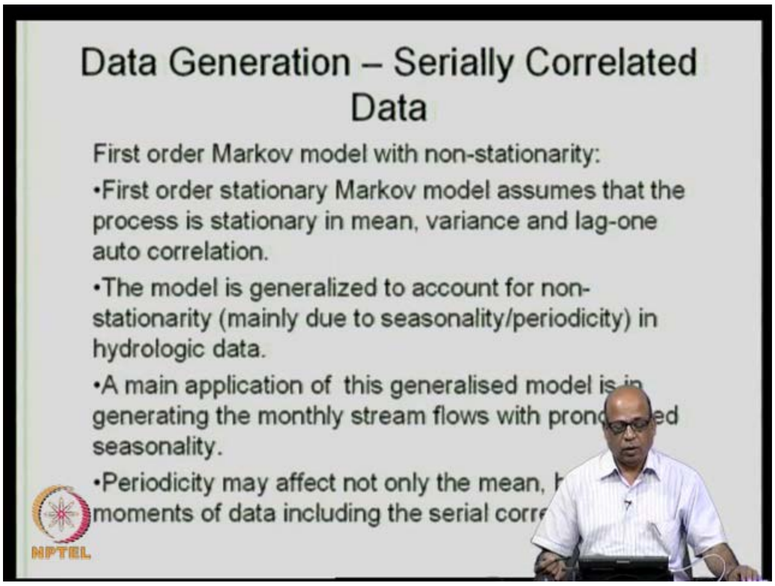
$$X_{j+1} = \mu_x + \rho_1\left(X_j - \mu_x\right) + t_{j+1}\sigma_x\sqrt{1-\rho_1^2}$$

Standard normal deviate

- Stationary w.r.t mean, variance and lag-one correlation
- Known sample estimates of μ_x, σ_x, ρ_1
- Assume X_1 (= μ_x)
- Generate values X_2, X_3, X_4, X_5……

So, we will now consider the first order Markov model with non stationarity. Now, the model that we just considered earlier, it was meant for annual flows. We will start relaxing the requirement that these need to be stationary and then what we do is, we write this model for seasonal flows. The seasons can be either months in which case, we will have twelve periods they can be monsoon, non monsoon season in which we will have two seasons or monsoon summer and winter three seasons. We May have we May consider ten day durations in which case we May have 36 or 37 time periods, according to how we (()). Like this, we will now consider the intra year periods over which we are interested in generating the flows.

(Refer Slide Time: 09:39)



Data Generation – Serially Correlated Data

First order Markov model with non-stationarity:

•First order stationary Markov model assumes that the process is stationary in mean, variance and lag-one auto correlation.

•The model is generalized to account for non-stationarity (mainly due to seasonality/periodicity) in hydrologic data.

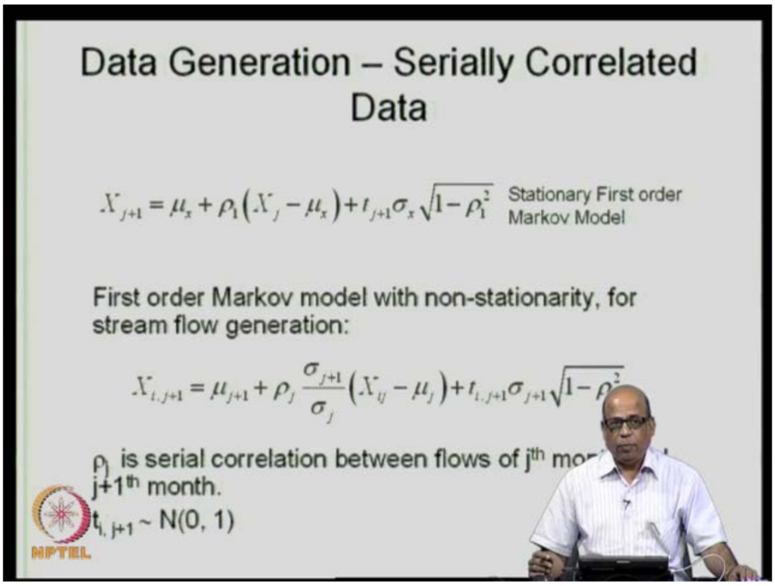
•A main application of this generalised model is in generating the monthly stream flows with pronc ed seasonality.

•Periodicity may affect not only the mean, h moments of data including the serial corr

So, the same model now we generalize to account for non stationarity and this non stationarity essentially arises from the periodicity, as I just mentioned the June month flow of this year May be correlated with the June month flow of the previous year. So, there May be a twelve month periodicity, there May be a six month periodicity, there May be two year periodicity and so on depending on the type of data that we consider. And, this kind of periodicity introduces non stationarity in the data and this non stationarity, we will build into this model now build into the first order Markov model.

A main application of the Markov model considering the non stationarity in the data is essentially for monthly stream flow generation, it has been very effectively used for generating monthly stream flows in situations, where there is a pronounced periodicity or seasonality. And, the periodicity as you know now will affect not only the mean and standard deviation, but also it will affect the lag one correlations, all of which appear in our Thomas fiering model or the Markov model, first order Markov model.

(Refer Slide Time: 11:02)



Data Generation – Serially Correlated Data

$$X_{j+1} = \mu_x + \rho_1 \left(X_j - \mu_x \right) + t_{j+1} \sigma_x \sqrt{1-\rho_1^2}$$ Stationary First order Markov Model

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

$$X_{i,j+1} = \mu_{j+1} + \rho_j \frac{\sigma_{j+1}}{\sigma_j} \left(X_{ij} - \mu_j \right) + t_{i,j+1} \sigma_{j+1} \sqrt{1-\rho^2}$$

ρ_j is serial correlation between flows of j^th mo
j+1^th month.
$t_{i,j+1} \sim N(0, 1)$

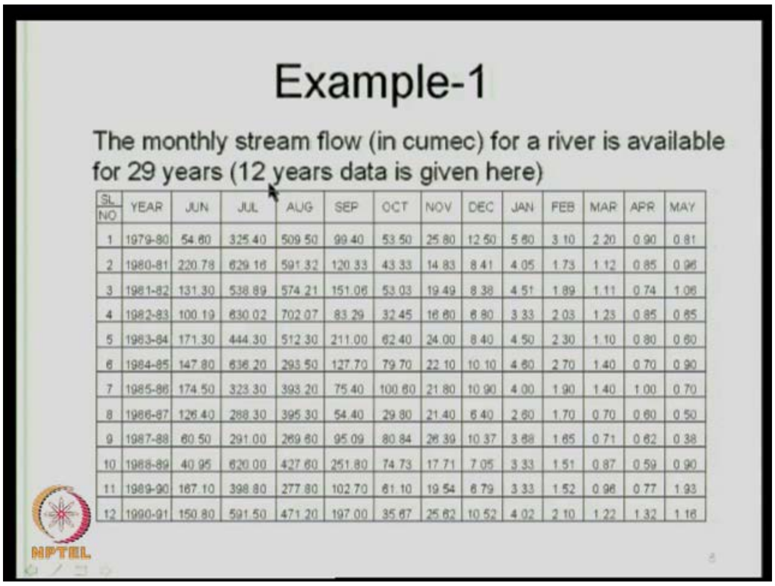
So, from your stationary model we start introducing the non stationarity in the mean lag one correlation and the standard deviation. By introducing one more index here so i is the year and j is the month. So, we are generating for the i-th year and j plus 1 at a month or j plus 1 at the season if you like. So, what was mu x which means you had one mean for the entire sequence, we now convert that into mu of that particular season for which you are generating the data. So, mu j plus 1 plus instead of calling it as rho 1 lag one correlation, we denote it by rho j where rho is in fact, the lag one correlation and rho j indicates the lag one correlation between the month j and the j plus 1, the month j and month j plus 1.

Because we are generating for the month j plus 1, the rho j will be the dependence of j plus 1 will indicate the dependence of j plus 1 th month's flow on its previous month's flow. So, rho j indicates the correlation between the flows of month j and the flows of month j plus 1, similarly sigma j indicates the standard deviation of the month j.

And, the random component we are introducing here for every value, you generate here the random component will be different. So, it will have the same indices as your flow has the generating; the flow to be generated has. So, X i j plus 1 you have t i j plus 1, where t i j plus 1 is the drawn from N 0 1; that means, this follows normal distribution, standard normal distribution. So, using this expression then we should be able to generate monthly flows from the historical available flows.

So, we would first estimate the moment's mu j plus 1, which is mu j for j is equal to 1 to m where m is a number of seasons that we are considering. If it is a monthly model, m will be equal to 12. So, for each of the months we have the mean standard deviations and the lag one correlations with the next month. So, we have the data ready we start with an assumed initial value, which is typically assumed to be the mean itself. So, that this term vanishes exactly the same way as we did for the annual flows.

(Refer Slide Time: 14:10)



Example-1

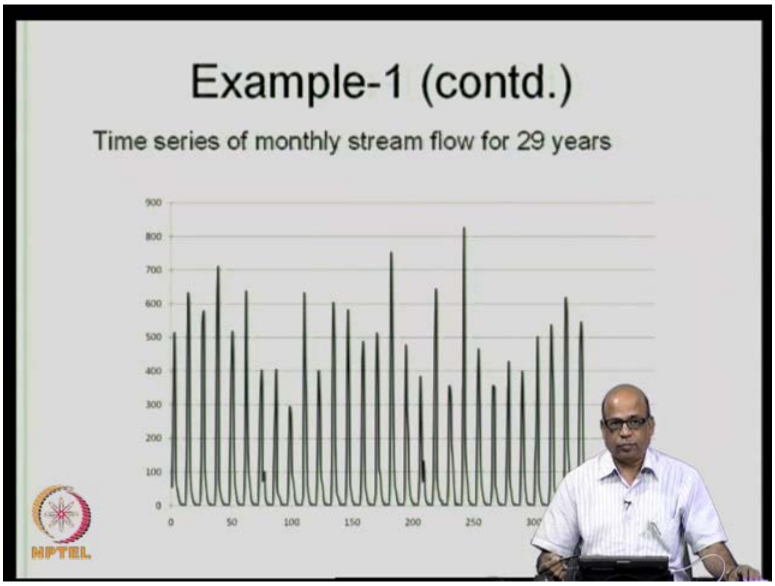
The monthly stream flow (in cumec) for a river is available for 29 years (12 years data is given here)

SL NO	YEAR	JUN	JUL	AUG	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY
1	1979-80	54.60	325.40	509.50	99.40	53.50	25.80	12.50	5.60	3.10	2.20	0.90	0.81
2	1980-81	220.78	629.16	591.32	120.33	43.33	14.83	8.41	4.05	1.73	1.12	0.85	0.96
3	1981-82	131.30	538.89	574.21	151.06	53.03	19.49	8.38	4.51	1.89	1.11	0.74	1.06
4	1982-83	100.19	630.02	702.07	83.29	32.45	16.60	6.80	3.33	2.03	1.23	0.85	0.65
5	1983-84	171.30	444.30	512.30	211.00	62.40	24.00	8.40	4.50	2.30	1.10	0.80	0.60
6	1984-85	147.80	636.20	293.50	127.70	79.70	22.10	10.10	4.60	2.70	1.40	0.70	0.90
7	1985-86	174.50	323.30	393.20	75.40	100.60	21.80	10.90	4.00	1.90	1.40	1.00	0.70
8	1986-87	126.40	288.30	395.30	54.40	29.80	21.40	6.40	2.60	1.70	0.70	0.60	0.50
9	1987-88	60.50	291.00	269.60	95.09	80.84	26.39	10.37	3.68	1.65	0.71	0.62	0.38
10	1988-89	40.95	620.00	427.60	251.80	74.73	17.71	7.05	3.33	1.51	0.87	0.59	0.90
11	1989-90	167.10	398.80	277.80	102.70	61.10	19.54	6.79	3.33	1.52	0.96	0.77	1.93
12	1990-91	150.80	591.50	471.20	197.00	35.67	25.62	10.52	4.02	2.10	1.22	1.32	1.16

We will consider an example now; we have stream flows at a particular river for 29 years, we have shown 12 years data here. So, like this we have for 29 years the data, the data is like this June, July etc up to May from 1979 80 goes on for 29 years. So, you have the data collected for 29 years. Remember, if you wanted to model it using a non stationary model what you will do, that you will compute the mean for each of these months, June month using the 29 values of June, July month using the 29 values of July and so on. So, you will have means for all the twelve months, similarly standard deviations.

Similarly, you take the pairs June July and get rho 1, July August rho 2 etc, then May and June has rho 12. So, when you are considering the last month the twelfth month, the correlation will be with respect to the next month, that the next following month which will be June. So, you will estimate all the parameters based on this data and use them in the model.

(Refer Slide Time: 15:36)

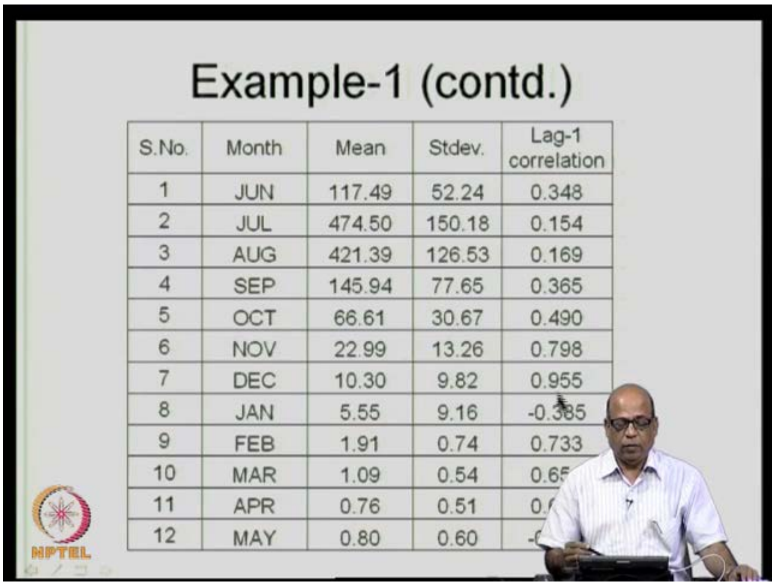


Example-1 (contd.)

Time series of monthly stream flow for 29 years

So, this is the time series for twenty years time series of the flows same data is shown in a figure here and from this data, we now compute the mean standard deviation and the lag one correlation.

(Refer Slide Time: 15:48)

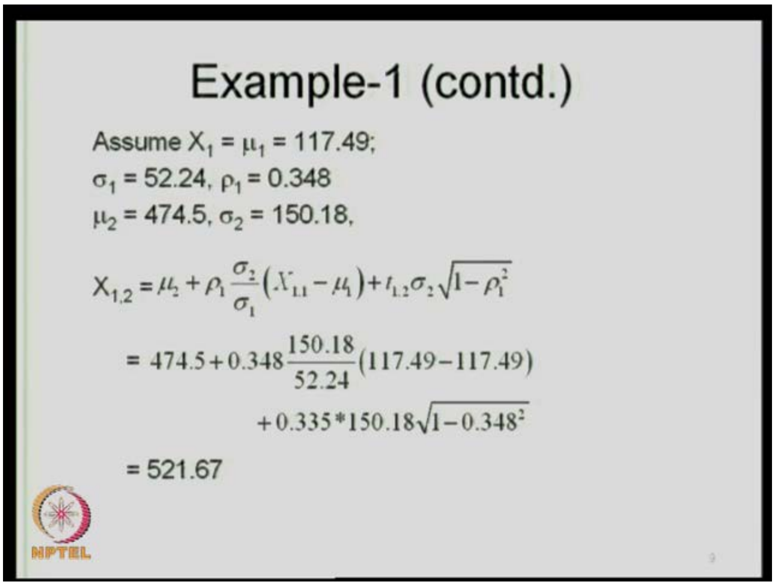


Example-1 (contd.)

S.No.	Month	Mean	Stdev.	Lag-1 correlation
1	JUN	117.49	52.24	0.348
2	JUL	474.50	150.18	0.154
3	AUG	421.39	126.53	0.169
4	SEP	145.94	77.65	0.365
5	OCT	66.61	30.67	0.490
6	NOV	22.99	13.26	0.798
7	DEC	10.30	9.82	0.955
8	JAN	5.55	9.16	-0.385
9	FEB	1.91	0.74	0.733
10	MAR	1.09	0.54	0.65
11	APR	0.76	0.51	0.
12	MAY	0.80	0.60	-

Now, these values here provide the mean standard deviation and lag one correlation. I repeat again the lag one correlation that we are writing here, e is with respect to the next month that is this indicates the lag one correlation between the flows of June to the flow of July. Similarly, this indicates the lag one correlation between the flows of May with flow of June.

(Refer Slide Time: 16:28)



Example-1 (contd.)

Assume $X_1 = \mu_1 = 117.49$;
$\sigma_1 = 52.24$, $\rho_1 = 0.348$
$\mu_2 = 474.5$, $\sigma_2 = 150.18$,

$$X_{1,2} = \mu_2 + \rho_1 \frac{\sigma_2}{\sigma_1}(X_{1,1} - \mu_1) + t_{1,2}\sigma_2\sqrt{1-\rho_1^2}$$

$$= 474.5 + 0.348\frac{150.18}{52.24}(117.49 - 117.49)$$

$$+ 0.335 * 150.18\sqrt{1-0.348^2}$$

$$= 521.67$$

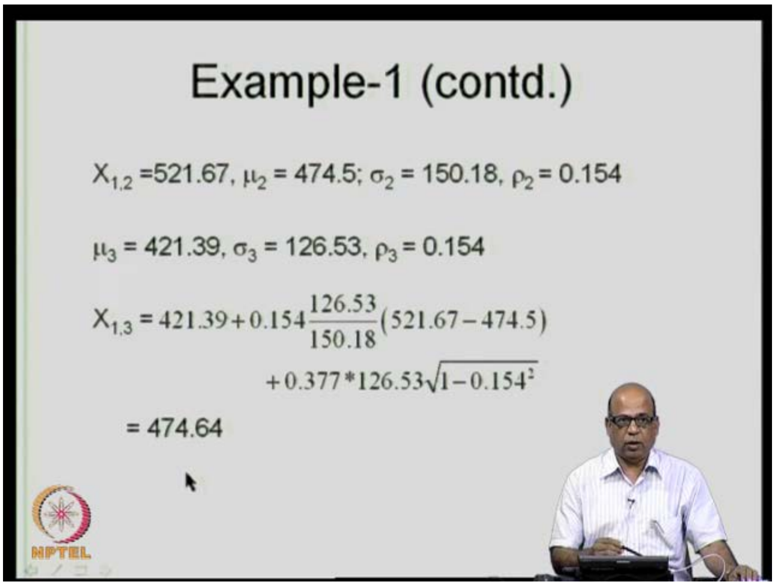
Once, we are ready with this we start generating the model generating the data, we assume the first value let’s say X 1 is assumed to be the same as the mean of the first month which is 117.49. And, sigma 1 is given here sigma 1 is 52.24 and you want to generate the second month’s flow using the first month flow. So, your correlation will be

0.348 similarly, because you are generating the second month flow you will get mu 2 and sigma 2 that, you will use from mu 2 is 474.5 and sigma 2 is 150.18.

These values we use and write X 1 2 what does this mean, the flow for the first year for the second month. So, you would have assumed X 1 1, here this is X 1 1, this is X 1 1 is equal to mu 1 is what we have assumed and starting with that, we will write this to be 474.5 plus 0.348 into 150.18 by 52.24 etc. I just want to explain one thing in the expression here, when we wrote from the stationary model, stationary first order model to a non stationary model, here for a second term we introduced the ratio sigma j plus 1 by sigma j, you can see here X j minus mu j by sigma j here this is nothing but the standardized value. So, we use the standardized value and similarly, you get here X i j plus 1 minus mu j plus 1 by sigma j plus 1.

So, that is why we introduce this ratio sigma j plus 1 by sigma j and that is what we are using here. So, this will be sigma j plus 1 by sigma j and we write this value to be assumed, because we are assuming X 1 1 to be mu 1 itself. So, this term goes this is a standard normal deviate, pick it up from the table or otherwise and this is your standard deviation for the second month in to root of 1 minus rho 1 square. So, you get 521.67.

(Refer Slide Time: 19:17)



Example-1 (contd.)

$X_{1,2}$ =521.67, μ_2 = 474.5; σ_2 = 150.18, ρ_2 = 0.154

μ_3 = 421.39, σ_3 = 126.53, ρ_3 = 0.154

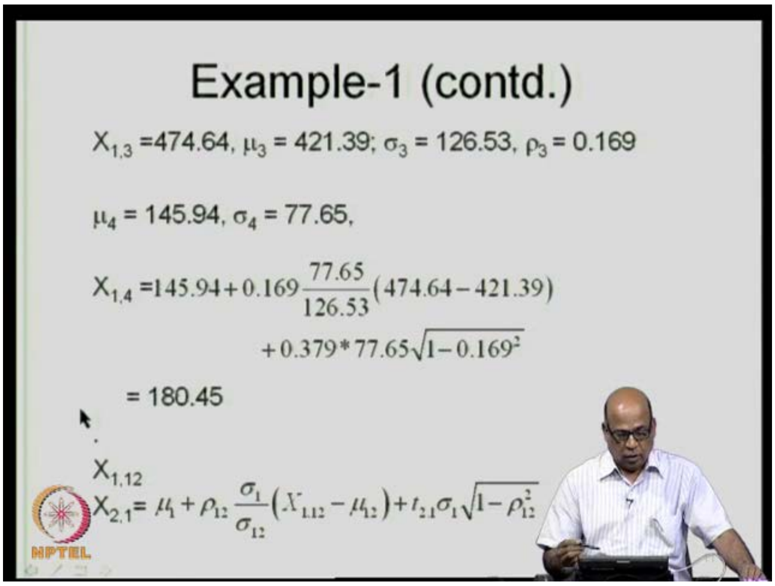
$$X_{1,3} = 421.39 + 0.154\frac{126.53}{150.18}(521.67 - 474.5) + 0.377 * 126.53\sqrt{1 - 0.154^2}$$

= 474.64

Now, we use this 521.67 to generate X 1 3 now we get X 1 3 for which, we will also require mu 3 sigma 3 and rho 2 in this particular case we need. So, we use these values and generate the next values here, there is another small mistake here this is not rho 3, we will just take. So, we get the value as 474.64 this was 0.348 that we used from here.

So, the next value that we will be using is 0.154. So, 0.154 is what we use this is not rho 3, but rho 2 here. So, rho 2 is what we will be using, because it is a correlation between month 2 and month 3. So, X 1 3 is what we are getting so, we generate it to be equal to 474.64.

(Refer Slide time: 20:28)



Example-1 (contd.)

$X_{1,3} = 474.64,\ \mu_3 = 421.39;\ \sigma_3 = 126.53,\ \rho_3 = 0.169$

$\mu_4 = 145.94,\ \sigma_4 = 77.65,$

$$X_{1,4} = 145.94 + 0.169\frac{77.65}{126.53}(474.64 - 421.39) + 0.379 * 77.65\sqrt{1 - 0.169^2}$$

$= 180.45$

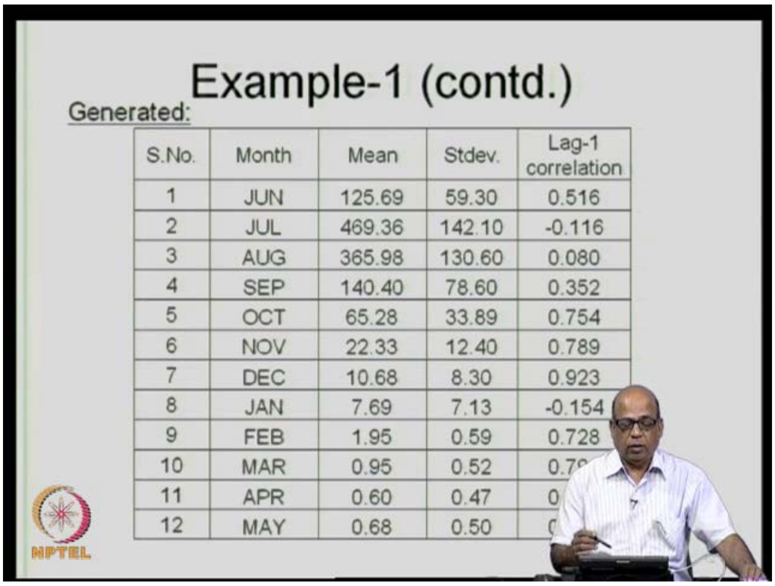
$X_{1,12}$

$$X_{2,1} = \mu_1 + \rho_{12}\frac{\sigma_1}{\sigma_{12}}(X_{1,12} - \mu_{12}) + t_{2,1}\sigma_1\sqrt{1 - \rho_{12}^2}$$

Like this, we keep on doing from the third value we generate the fourth value and so on. This is a monthly model, so we generate from first month, second month etc up to twelfth month, like this we keep on going to twelfth month. Once, we reach the twelfth month we go to the next year flows. So, we write this as 2 1 second year first flow, first month's flow and what will be using there mu 1 plus rho 12, because rho 12 is what drives the first month's flow, the dependence of the first month's flow on the previous month's flow which is the twelfth month flow, that is what is given by rho 12.

We write the expression for the second year first month flow, then we carry on second year second month flow etc like this. Like this, you generate for 50 years, 100 years, 150 years etc depending on the need. So, like this we proceed and generate the time series of monthly flows.

(Refer Slide Time: 21:31)



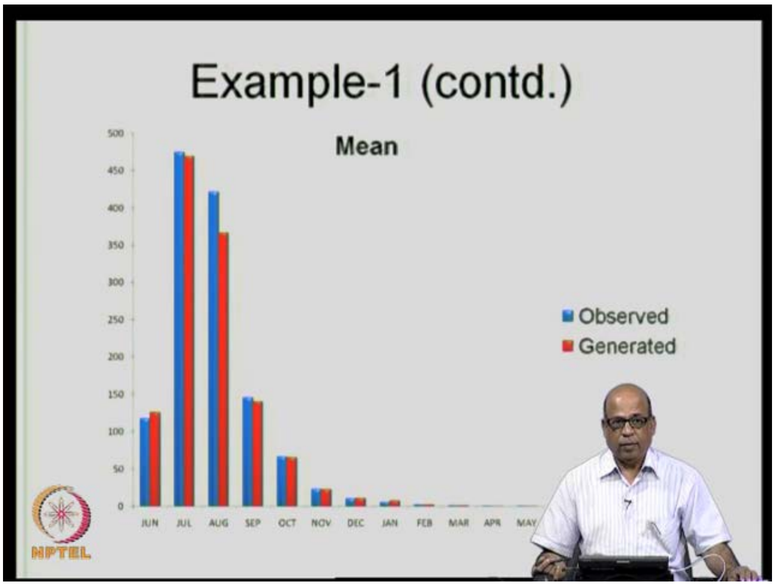
Example-1 (contd.)

Generated:

S.No.	Month	Mean	Stdev.	Lag-1 correlation
1	JUN	125.69	59.30	0.516
2	JUL	469.36	142.10	-0.116
3	AUG	365.98	130.60	0.080
4	SEP	140.40	78.60	0.352
5	OCT	65.28	33.89	0.754
6	NOV	22.33	12.40	0.789
7	DEC	10.68	8.30	0.923
8	JAN	7.69	7.13	-0.154
9	FEB	1.95	0.59	0.728
10	MAR	0.95	0.52	0.7
11	APR	0.60	0.47	0
12	MAY	0.68	0.50	0

Let’s consider another example, what we then do is like this we generate for 50 years and then compare; in this particular case we generated it for hundred years like this 2 1 up to 212, 31 312 etc like this it keeps on going up to 100 1 100 12, 100 comma 12 that is a hundredth year all the twelve months. We generate the values and then compute the mean standard deviation and the lag one correlation of the generated data. Typically, when we do this we discard the first few values and compute the mean standard deviation and lag one correlation of the generated data. Now, these should be approximately the same as the mean standard deviation and lag one correlation of the historical data.

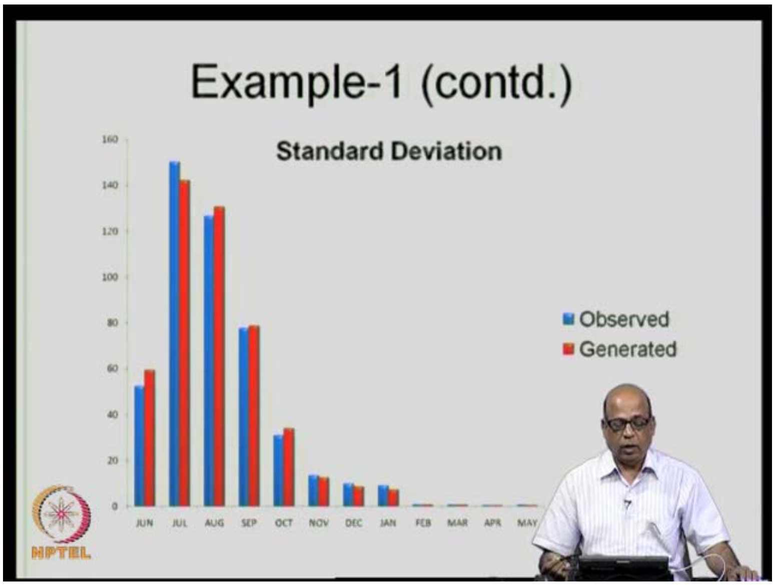
(Refer Slide Time: 22:36)



Example-1 (contd.)

So, this we compare by drawing bar charts. So, this is a general procedure that you generate the data for a sufficiently long period of time, for a sufficiently long sequence of data you generate. Discard the first few values and then using the remaining data, you compute the mean standard deviation and lag one correlation and compare these with the historical that is the observed so, the observed is historical data and the generated data. Typically, in this particular example the means match fairly well so this is an observed data, this is a generated data.

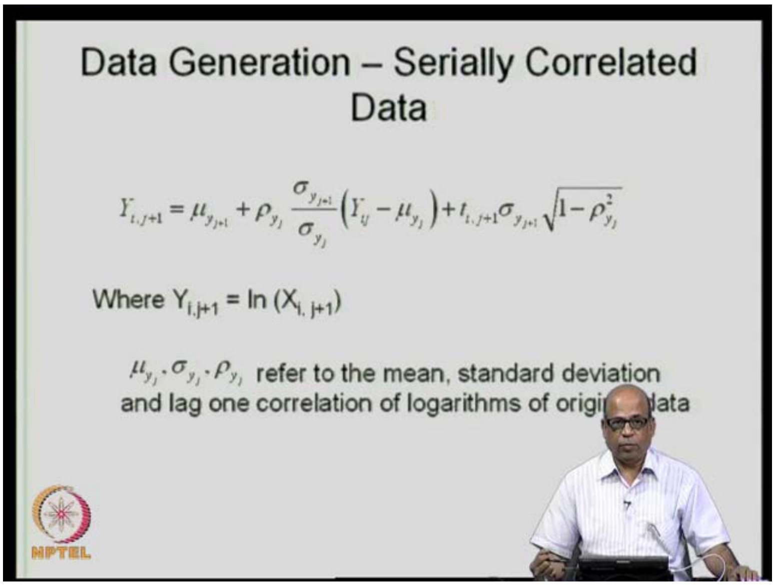
(Refer Slide Time: 23:16)



Example-1 (contd.)

Similarly, the standard deviation this is the observed data, this is a generated data. So, mean and standard deviation compare fairly well.

(Refer Slide Time: 23:26)



Data Generation – Serially Correlated Data

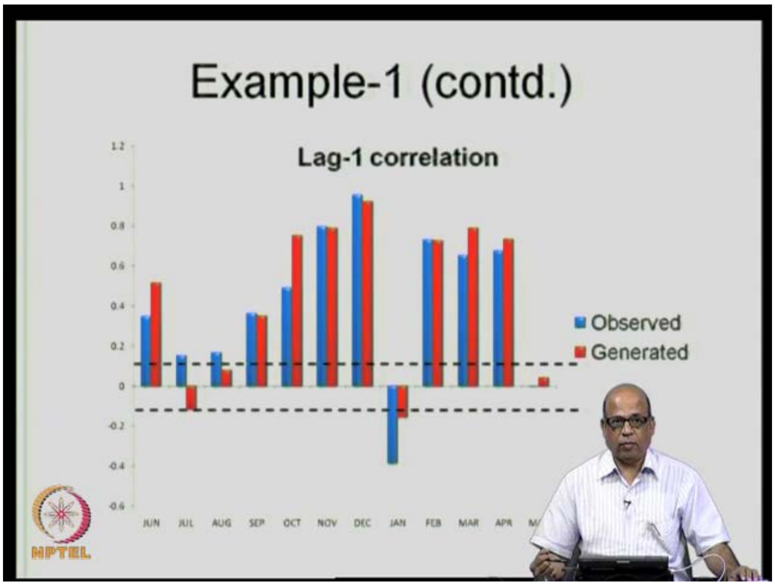
$$Y_{i,j+1} = \mu_{y_{j+1}} + \rho_{y_j}\frac{\sigma_{y_{j+1}}}{\sigma_{y_j}}\left(Y_{ij} - \mu_{y_j}\right) + t_{i,j+1}\sigma_{y_{j+1}}\sqrt{1-\rho_{y_j}^2}$$

Where $Y_{i,j+1} = \ln(X_{i,j+1})$

$\mu_{y_j}, \sigma_{y_j}, \rho_{y_j}$ refer to the mean, standard deviation and lag one correlation of logarithms of origi data

So, let’s see what happens to the lag one correlation.

(Refer Slide Time: 23:27)



Example-1 (contd.)

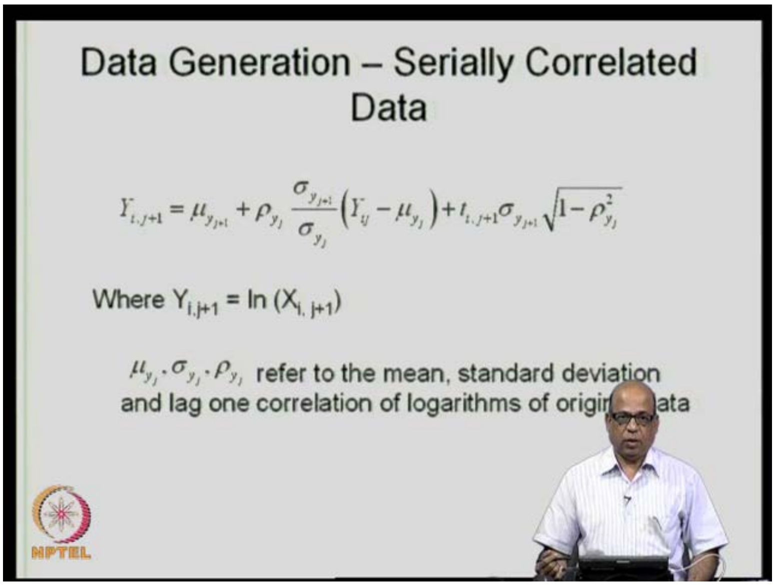
So, similarly for the lag one correlation we had observed values and we also have the generated values. Sometimes, the lag one correlations do not perform really well or specifically for example, here you had a lag one correlation of order of about minus point just around minus 0.2 whereas, generated one has reached almost minus 0.4, if this repeats many times within a twelve months period, then you should be concerned. In this particular case, you also see that at one point the lag one correlation was on the positive side whereas, the generated one is on the negative side.

As long as both of them are insignificant, statistically insignificant then you do not have to worry much whether, it is on the positive side or negative side, because anyway it is insignificant. But if they are significant lag one correlations, statistically significant lag one correlations and they show different signs, then you should be concerned about the generated data. This is a significance band here. So, this is slightly about the significance band whereas, this is the insignificance band. So, there is a cause for concern here that the model is not really performing well in terms of lag one correlation whereas; in terms of the mean and the standard deviation the model is fairly acceptable.

So, if you have applications where your lag one serial preservation of the lag one serial correlation is extremely important and you do not want to sacrifice on these two months. For example, you are talking about the month July here and the month January, if the lag one correlations have to be preserved for your application for these months, then you may have to start looking at other possibilities or looking at improvement of this model and so on. Otherwise, this is a fairly acceptable situation.

We will also have situations, where the original data as shown here may not perfectly fit or may not be acceptable for the assumption of normal distribution and then you may get some unacceptable results, in terms of the mean or standard deviation or lag one correlations in as much as they do not compare well with the observed or the generated values do not compare well with the observed data.

(Refer Slide Time: 26:38)



Data Generation – Serially Correlated Data

$$Y_{i,j+1} = \mu_{y_{j+1}} + \rho_{y_j} \frac{\sigma_{y_{j+1}}}{\sigma_{y_j}} \left(Y_{ij} - \mu_{y_j} \right) + t_{i,j+1} \sigma_{y_{j+1}} \sqrt{1-\rho_{y_j}^2}$$

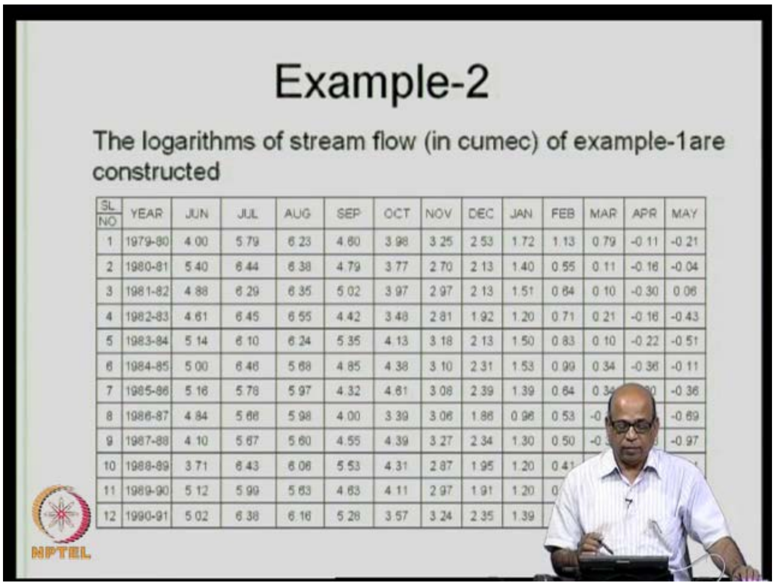
Where $Y_{i,j+1} = \ln(X_{i,j+1})$

$\mu_{y_j} \cdot \sigma_{y_j} \cdot \rho_{y_j}$ refer to the mean, standard deviation and lag one correlation of logarithms of origi ata

Then, what you must try is try with the logarithms of the flows; that means, your original data may not be normal distribution, but it is possible that the logarithms of the flows may be normally distributed, which means that the flows can be approximated as log normal distributions. In which case, we write the same model exactly the same model, but in terms of the logarithms of the flows. So, we use the transformation Y i j plus 1 is equal to logarithm of X i j plus 1. Simply convert the flows into logarithms and then write the same model in terms of the logarithms of the flow. Remember here, mu Y j plus 1 is with respect to the logarithm. So, all these moments here mu y j, sigma y j and rho y j-these refer to the mean, respectively mean standard deviation and lag one correlation of logarithms of original data.

So, if you apply the model with the original data and you find that the comparison of the generated data with the observed data is not acceptable, then you may try with the logarithm’s data, logarithmic flows log transformed flows the same model, but now written in terms of the log transformed flows.

(Refer Slide Time: 28:03)



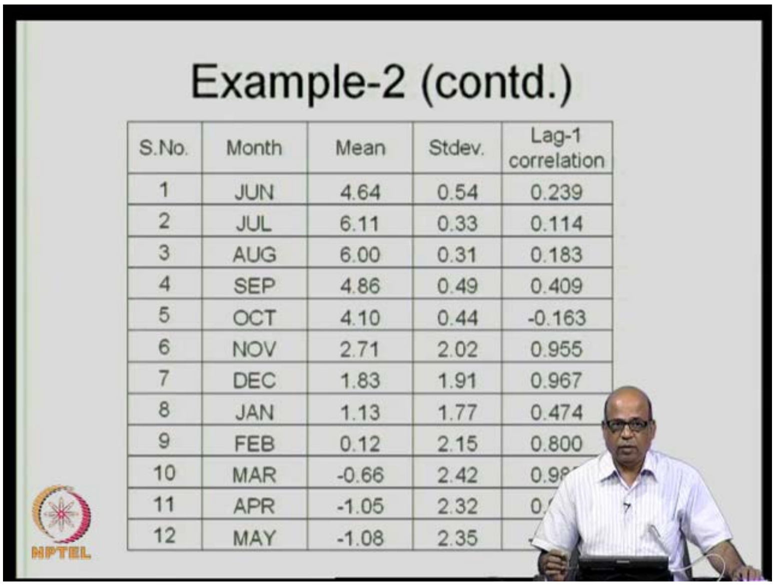
Example-2

The logarithms of stream flow (in cumec) of example-1are constructed

SL NO	YEAR	JUN	JUL	AUG	SEP	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY
1	1979-80	4.00	5.79	6.23	4.60	3.98	3.25	2.53	1.72	1.13	0.79	-0.11	-0.21
2	1980-81	5.40	6.44	6.38	4.79	3.77	2.70	2.13	1.40	0.55	0.11	-0.16	-0.04
3	1981-82	4.88	6.29	6.35	5.02	3.97	2.97	2.13	1.51	0.64	0.10	-0.30	0.06
4	1982-83	4.61	6.45	6.55	4.42	3.48	2.81	1.92	1.20	0.71	0.21	-0.16	-0.43
5	1983-84	5.14	6.10	6.24	5.35	4.13	3.18	2.13	1.50	0.83	0.10	-0.22	-0.51
6	1984-85	5.00	6.46	5.68	4.85	4.38	3.10	2.31	1.53	0.99	0.34	-0.36	-0.11
7	1985-86	5.16	5.78	5.97	4.32	4.61	3.08	2.39	1.39	0.64	0.34	[illegible]	-0.36
8	1986-87	4.84	5.66	5.98	4.00	3.39	3.06	1.86	0.96	0.53	[illegible]	[illegible]	-0.69
9	1987-88	4.10	5.67	5.60	4.55	4.39	3.27	2.34	1.30	0.50	[illegible]	[illegible]	-0.97
10	1988-89	3.71	6.43	6.06	5.53	4.31	2.87	1.95	1.20	0.41	[illegible]	[illegible]	[illegible]
11	1989-90	5.12	5.99	5.63	4.63	4.11	2.97	1.91	1.20	[illegible]	[illegible]	[illegible]	[illegible]
12	1990-91	5.02	6.38	6.16	5.28	3.57	3.24	2.35	1.39	[illegible]	[illegible]	[illegible]	[illegible]

Let’s, examine the data that was given in the previous example the same data, but we will convert that into logarithm of the flows. So, again 12 years are shown, but we have 29 years of data for that. So, you convert that into logarithm of flows. When you are converting into logarithm of flows generally, you face a difficulty that some of the flows may be zero, in which case log 0 is not defined and therefore, you may face that difficulty. What we generally do is, if you have 0 values put it to a very small value let’s say 0.005 or some such thing.

(Refer Slide Time: 28:59)



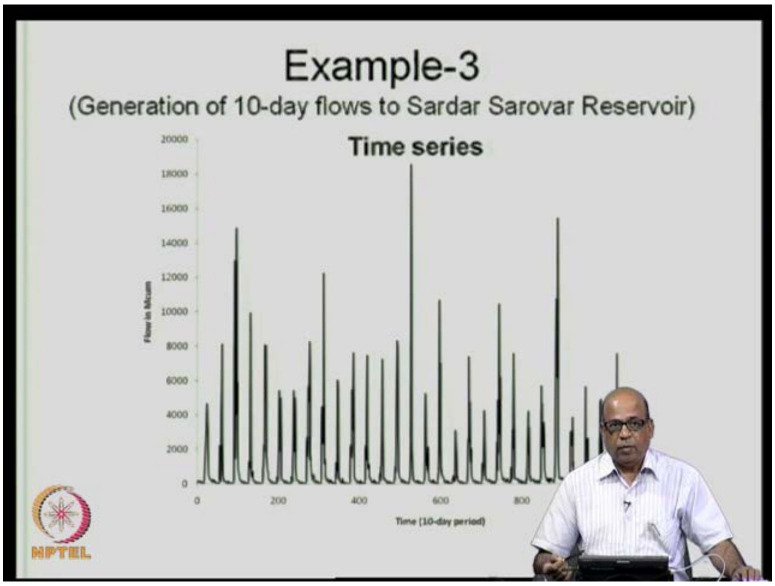
Example-2 (contd.)

S.No.	Month	Mean	Stdev.	Lag-1 correlation
1	JUN	4.64	0.54	0.239
2	JUL	6.11	0.33	0.114
3	AUG	6.00	0.31	0.183
4	SEP	4.86	0.49	0.409
5	OCT	4.10	0.44	-0.163
6	NOV	2.71	2.02	0.955
7	DEC	1.83	1.91	0.967
8	JAN	1.13	1.77	0.474
9	FEB	0.12	2.15	0.800
10	MAR	-0.66	2.42	[illegible]
11	APR	-1.05	2.32	[illegible]
12	MAY	-1.08	2.35	[illegible]

Very small value compared to the other values that are there in the series. So, that you can use the log transformation and because you are talking about log transformation if you have values less than one, and then you may also get a negative value here, that is perfectly fine. So, we use the log transformed values and then we get the mean standard deviation and lag one correlation associated with the log transformed values.

We use these in the model that we just defined and generate values of Y, remember we are generating values of log transformed values now log transformed data. When we do that and compare the mean standard deviation etc, this is how it looks the standard deviations appear like this and the lag one correlations appear like this. The lag one correlations again there are some months, in which they do not seem to tally well, they do not seem to compare well, but as long as they are statistically insignificant and it is acceptable for those particular months for the particular applications that you are talking about, then it is fine.

(Refer Slide Time: 30:09)



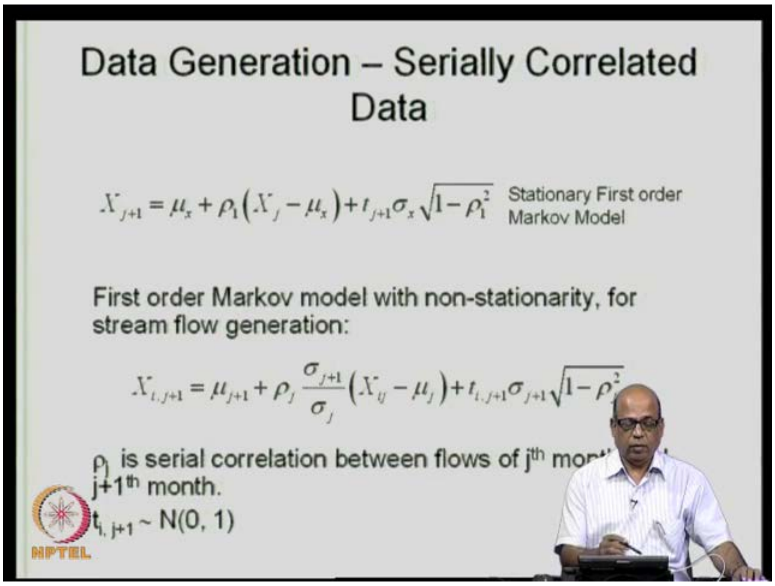
Example-3

(Generation of 10-day flows to Sardar Sarovar Reservoir)

With this methodology, now we will demonstrate one shorter time period flow generation as I said, this is a seasonal Thomas fiering model where the seasons can be monsoon, non monsoon, summer etc or the seasons can be months January, February etc or in many applications especially for irrigation reservoirs, hydro power reservoirs etc. You may be talking about other durations for example, ten daily duration, weekly duration and in certain cases daily duration and so on. But when we are applying the Thomas fiering model or the first order Markov model with the assumption of normal distribution for the flows, we must be alert to the situation that as you start reducing your time duration. The assumption of normal distribution for the flows may not be strictly valid and therefore, you can should not use this model for very small duration like daily flows, six days flows and so on.

Often, it has been used successfully for ten day flows and we demonstrate one such example here ten day flows to Sardar Sarovar Reservoir in India. This is the data, this data is available for fairly long time, this is a ten day time period. So, what we do is a year is divided into 36 time periods, 36 time intervals during a year. Like in the case of monthly time period, you had 12 intervals you had 36 time intervals during a year and this shows the why so, this shows the flow in million cubic meters. So, this is just a time series plot.

(Refer Slide Time: 32:27)



Data Generation – Serially Correlated Data

$$X_{j+1} = \mu_x + \rho_1 \left(X_j - \mu_x \right) + t_{j+1} \sigma_x \sqrt{1 - \rho_1^2}$$ Stationary First order Markov Model

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

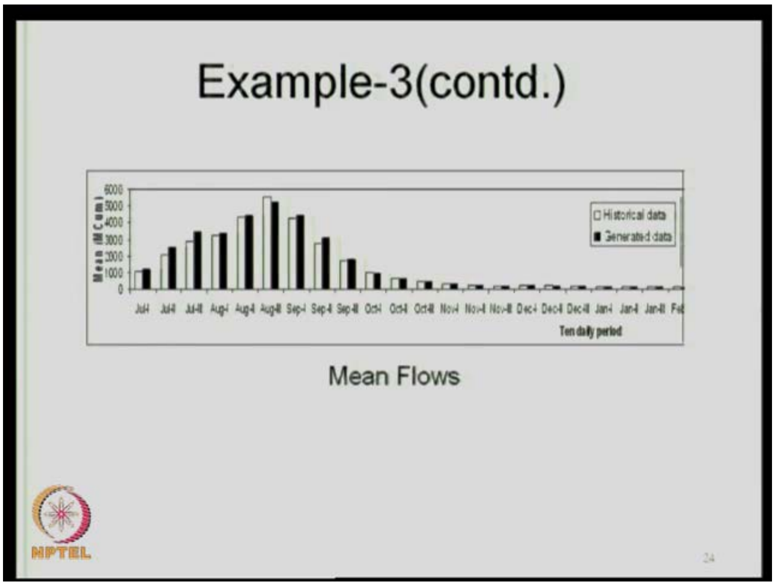
$$X_{i,j+1} = \mu_{j+1} + \rho_j \frac{\sigma_{j+1}}{\sigma_j} \left(X_{ij} - \mu_j \right) + t_{i,j+1} \sigma_{j+1} \sqrt{1 - \rho_j^2}$$

ρ_j is serial correlation between flows of j^{th} mon
$j+1^{th}$ month.

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

Now, with this time series plot then we use the Thomas fiering model as I just wrote here just for completeness, we will just see this is the Thomas fiering model or the first order Markov model, that we use i is the year and j is the month season and in this particular case day varies from 1 to 36. We have 36 time periods. So, for all these 36 time periods we would have computed the mean, the standard duration and the lag one correlations. We use this for 36 time periods and obtain the generated data and compare the generated data with the observed data.

(Refer Slide Time: 33:04)

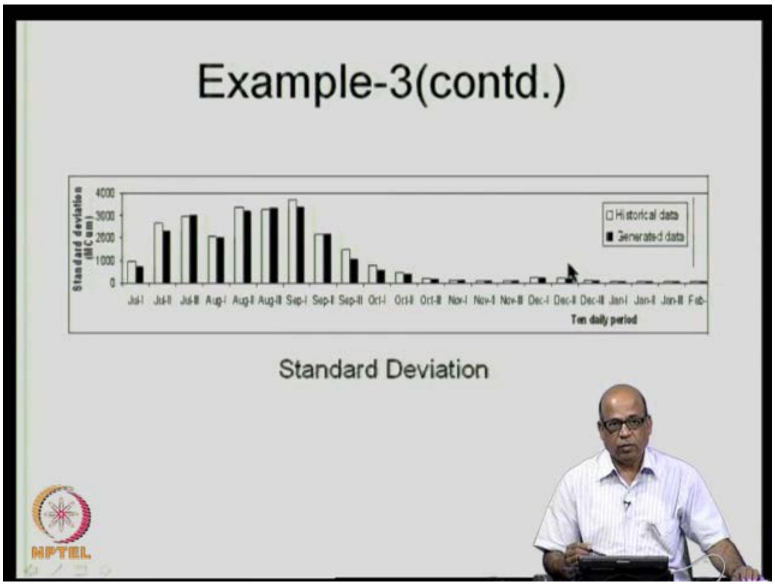


Example-3(contd.)

Mean Flows

So, when we do that the generated and the mean flows with generated and historical data we compare like this; this keeps going as it is mentioned here July 1, July 2, July 3 is the first month which is July in this case, first month three time periods July 1 to 10, 11 to 20 and so on. Similarly, August 1, August 2, August 3 are the three 10 day time periods in the month of August and so on. Like this we compare this figure extends, but because of lack of space I shown it here up to February first 10 day time period.

(Refer Slide Time: 33:54)

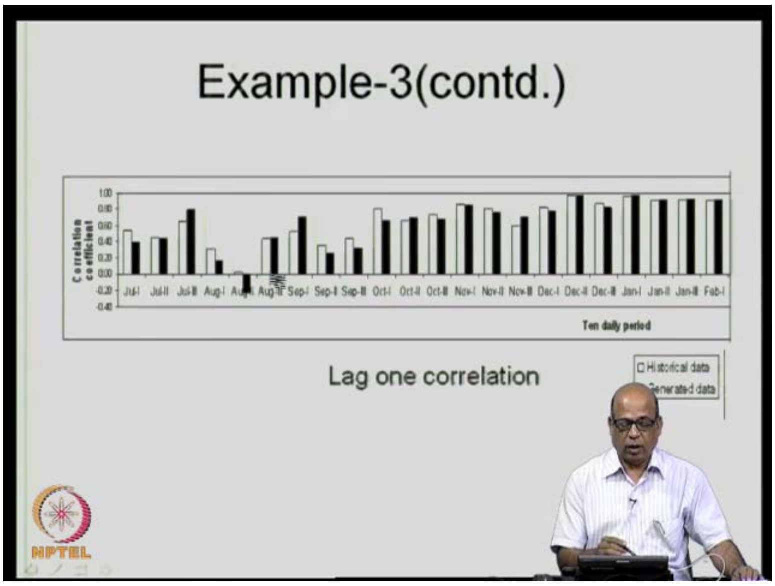


Example-3(contd.)

Standard Deviation

So, we compare the mean flow which is fairly acceptable in this case, similarly standard deviation we compare and then we compare the lag one correlations.

(Refer Slide Time: 33:59)



Example-3(contd.)

Lag one correlation

Lag one correlation again there is a problem with August second time period, but as long as this is insignificant then we can take it as acceptable, statically insignificant or we can take it as acceptable. So, this type of generation we use for in applications such as reservoir operation, what we exactly did in this particular case is that, the flows into Sardar Sarovar Reservoir from the historical available data. Let’s say, we had 40 years of data at the reservoir. Ten day periods, we generate this for let’s say 100 years, 200 years etc several such sequences, how do we generate several such sequences by using different sequences of the random numbers, that appear in the model. By generating different sequences of random numbers, we generate different sequences of flows.

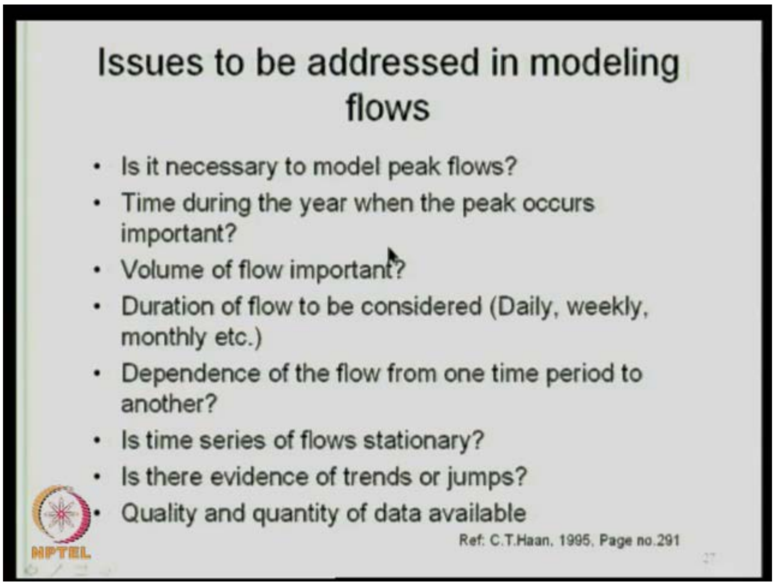
Like this, let’s say you have several sequences of 50 year data ten day period data. We use these data in the simulation of reservoir operation and generate let’s say, you are operating the reservoir for hydro power. We use these sequences of data in the simulation and then generate several levels of hydro power, as resulting from this particular in flow sequence and then start talking about, how the system performs for this level of generated data.

Because, the generated data has the same statistical properties as the observed data, instead of dealing with just one sequence of the observed data. We now, deal with several such sequences of observed data, so this is one direct application. In subsequent lectures, we will also see several other applications. So, when we are doing this a stochastic model of flows, there are several issues that we need to consider. That is essentially, you have an observed sequence of data from the observed sequence of data you want to generate another sequence or several such sequences of the data of the stream flow data, let's say which have the same statistical properties as your observed data.

The model that we just introduced namely, the first order stationary or non stationary Markov model. Suppose, you use this model and observe that your peak flows that where present in your historical data are not reproduced well, then you cannot may blame the model. Because, the model is not essentially meant to generate the peak flows, the model is meant to generate overall it has to it preserves the historical mean, it preserves the historical standard deviations and it preserves the lag one correlations.

So, there is no feature built in to the model to preserve the peak flows, let's say the annual maximum flows or the annual minimum flows. So, either way the either the maximum or the minimum if you are interested in those things, then you should not use a models like this. So, when you want to select a stochastic model, the purpose for which you want to use the model is extremely important.

(Refer Slide Time: 38:17)



Issues to be addressed in modeling flows

- Is it necessary to model peak flows?
- Time during the year when the peak occurs important?
- Volume of flow important?
- Duration of flow to be considered (Daily, weekly, monthly etc.)
- Dependence of the flow from one time period to another?
- Is time series of flows stationary?
- Is there evidence of trends or jumps?
- Quality and quantity of data available

Ref: C.T.Haan, 1995, Page no.291

And therefore, we address several issues before we go into the model selection itself, what type of model that you want. The first issue that we address in this is, is it necessary to model peak flows. If you want to model the peak flows, then you have to adopt those particular models which will preserve the peak flows. Let’s say you are talking about the stream flows to a reservoir for the purpose of reservoir operation for irrigation, for hydro power, for municipal water supply and so on.

In such situations you are not really concerned about preserving the peak flow themselves. So, you would be interested in on an average how the system behaves, in which case you are not really interested in the peak floods or the other extreme of droughts and so on. So, we are interested in looking at on an average how the system performs. So, the peak flows are not important. In situations, where you would like to have the peak flows also modeled; the next level question that you would like to ask is, is it important that you also consider the time during which the peak flow occurs. For example, the maximum flows may have occurred during the month of August, whether this fact is also important to be built into the model; that is the time during which the peak flow occurs is also to be built into the model.

Then, are we also interested in the volume of flow or just the fact that the flow has exceeded a particular threshold is enough, especially when we are modeling the peak flows. So, is the volume of the flow is important? Then is the duration of flow to be considered important? That is what I mean by that is that whether you want to model this for daily flows, weekly flows, monthly flows and so on. As it is very obvious, depending on the duration of the flow that you want to consider the type of the model that you would like to fit can be quite different. From annual to season to about months, you can still use the first order Markov model either stationary version or the non stationary version of it.

But as you start reducing your time periods, let’s say you come down to weekly time period, daily time period and so on. Then, the assumption of normality of flows will be violated and therefore, you may not be able to use such model then, you will have to go for different type of models, which will subsequently discuss in this course. Then, we will also look at dependence of the flow from one time period to another time period important in the Markov model, what did we do we introduced the lag one correlation coefficient. So, this dependence if it is important then we may need to introduce the correlation coefficients, it need not be only lag one correlation; it can be correlation with respect to let’s say twelve months behind that is the flow during a particular month, let’s say June of this year.

Its dependence on the flow during June month of previous year, which means we may considering lag of the order of twelve - lag of twelve. So, we need to understand the particular issue that is important for the specific application that we have in mind and also the structure of the data itself. If there is a significant correlation or a significant dependence of a particular months flow on another months flow in the with a certain lag period, then that correlation has to be built in to the particular model.

Then another important question that, we need to address affront is whether the time series is stationary whether, the data that we have is stationary, there are ways of assessing or estimating whether, the time series is stationary which will discuss subsequently. But, this is a very vital question, very critical question that we need to address, because if the time series is non stationary then we need to adopt (()) address a non stationarity as we just did first order non stationary Markov model, in which we wanted to build in the non stationarity due to the flows having different moments during different time periods specifically months.

Then, in the data is there in evidence of a jumps or trends whether, the data is it shows the continuously increasing trend, a continuously decreasing trend or is there a significant jump it was operating at a certain level and suddenly there is a jump and then it starts operating at a different level. So, is there an evidence of trends or jumps in the model? A most critical requirement or the issue that we need to be aware of, Is quality and quantity of data available itself. We may have for example, we may have flows at a particular location for last twelve to thirteen years, the quantity of data for any meaningful model will be quite small, if you have only 12 to 15 years of data.

So, typically for the models to be useful if you have at least about 30 years of data, then you can rely on the results that you get out of the data. Although, when these models were developed actually somewhere around 60’s, 70’s etc, these models have been used for data of lengths of as small as 15 years, 12 years and so on.

But, we expect that you have about 30 years of data to get meaningful results out of such models. And also, the quality of the data itself in many situations you may have data, but just by looking at the data you can say that this is of a very poor quality. In the sense that there may be missing data or there may be repetitions of the data which obviously, points to errors in the data and are hazy data in the sense that there may be values that are repeating, and there may be values that are missing in the sum sequence which indicates that the data has not been collected with reliable sources and so on.

So, we must be alert to situations where the data is of poor quality. So, quantity as well as quality of data is important. So, what we have listed here are the issues that we need to be alert to before we actually choose a particular stochastic model. Now, we progress on to another important topic, which also leads to development of stochastic models. So, far what we were doing we were expressing the time series X t as a deterministic component plus a stochastic component; X t is equal to d t plus epsilon t or e t we wrote earlier and all this analysis that we were doing was all on time domain in the time domain.

So, this was called this is called as analysis of data in the time domain. There is another elegant way of doing this is to convert the data and time domain into frequency domain. So, we write the time series in terms of frequencies specifically sin waves, cosine waves of varying frequencies and then we start looking at the time series in the frequency domain. So, that is called as analysis in the frequency domain.

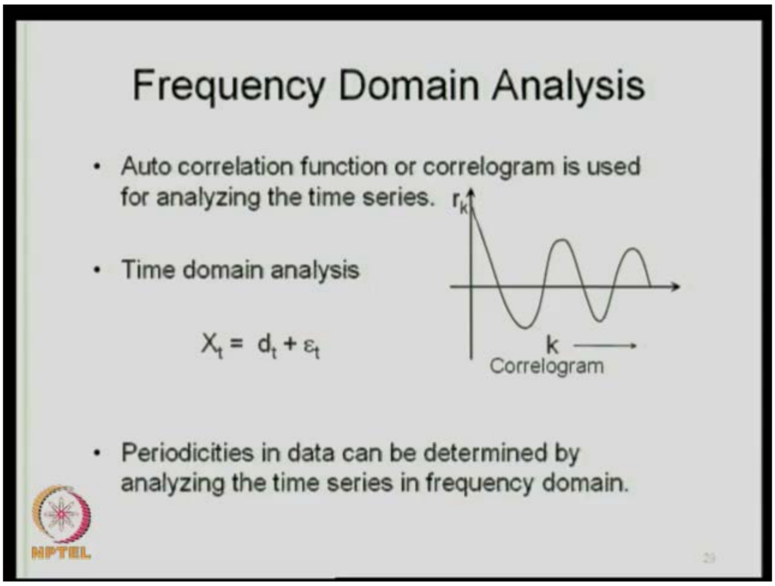
So, in today's class I will just introduce what we mean by analysis in the frequency domain, the details of that we will discuss in the next lecture. So, before going into the analysis in the frequency domain let us recapitulate, what we did in the analysis in the time domain. In the time domain, we expressed X t to be consisting of X t is the time series consisting of a deterministic component d t and a stochastic component epsilon t.

And it was our aim to capture what are these various components - what are these two components d t and epsilon t, if you recall the deterministic component d t can be either a trend or it can be a long term mean, around which there is a stochastic perturbations or there may be a periodicity. So, the data may be exhibiting a specific periodicity or there may be a jump or a drop, now these are the deterministic components.

Then the stochastic component, we need to capture the essence of the stochastic component and build it into the model. So, this is essentially the principle of analysis in the time domain. In the frequency domain, what we do is the X t which is the time series we write it as consisting of combination of several frequencies sin and cosine waves of several different frequencies and then start looking at the periodicities that come out significant, come out as significant periodicities inherited the data.

The frequency domain analysis is essentially used in hydrologic applications to capture the significant periodicities, as you recall in the when we discuss the correlograms or the auto correlations, the auto correlations also give an indication of whether the process is periodic or not, but it does not exactly indicate whether the periodicities that are indicated by the auto correlation function or in fact significant or not. Along with the auto correlation function, if we also use analysis in the frequency domain then we will be able to pin point the particular specific periodicities, that need to be considered in our models, in our generation models or the forecasting models and so on.

(Refer Slide Time: 50:31)



Frequency Domain Analysis

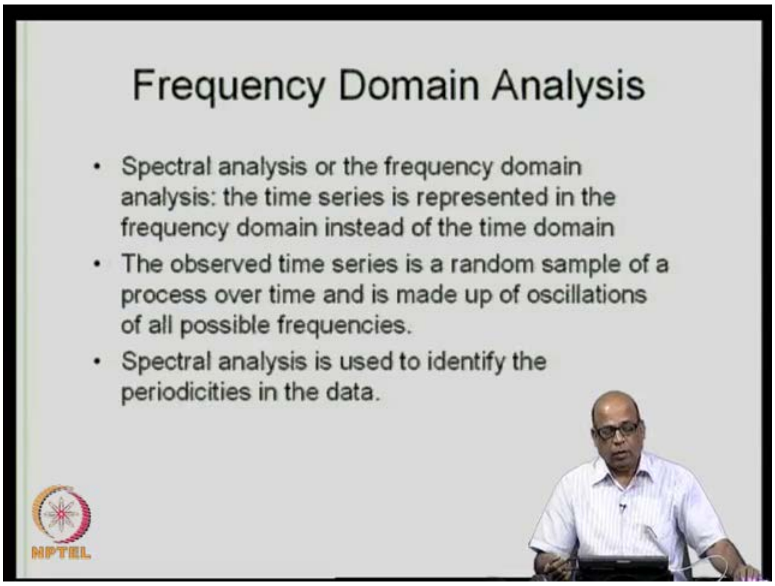
- Auto correlation function or correlogram is used for analyzing the time series.
- Time domain analysis

$$X_t = d_t + \varepsilon_t$$

- Periodicities in data can be determined by analyzing the time series in frequency domain.

So, in the frequency domain analysis this is just what I just mentioned that you may have a correlogram which is exhibiting a certain degree of periodicity here. So, if you have a periodic process the correlogram can be will appear something like this. Specifically, if you have monthly flows, the correlograms can be like this and then slowly it shows a decay. There will be a slow decay of the correlogram itself as I said in the time domain, we write X t is equal to d t plus e t and then we capture the deterministic component d t. So, periodicities in data can be determined by analyzing the time series in the frequency domain.

(Refer Slide Time: 51:31)

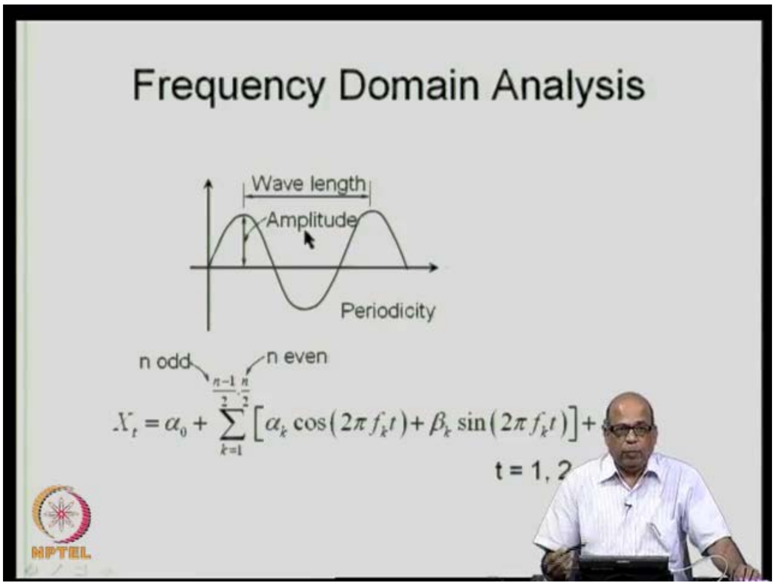


Frequency Domain Analysis

- Spectral analysis or the frequency domain analysis: the time series is represented in the frequency domain instead of the time domain
- The observed time series is a random sample of a process over time and is made up of oscillations of all possible frequencies.
- Spectral analysis is used to identify the periodicities in the data.

So, along with the information that you generate for using the correlogram you also generate information using analysis in the frequency domain to capture those periodicities. So, the frequency domain analysis which is also called as a spectral analysis in this the time series is represented in the frequency domain, instead of in the time domain. The observed series essentially the underlying principle here is as the observed time series is a random sample of a process over time, which is made up of oscillations of all possible frequencies. So, we convert the time domain the time information into frequency information and as I just mentioned the spectral analysis or the frequency domain analysis is used to identify the frequencies or periodicities inherent in the data.

(Refer Slide Time: 52:25)



Frequency Domain Analysis

n odd n even

$$X_t = \alpha_0 + \sum_{k=1}^{\frac{n-1}{2},\frac{n}{2}} \left[\alpha_k \cos(2\pi f_k t) + \beta_k \sin(2\pi f_k t)\right] +$$

t = 1, 2

Just a preliminary recap of what we mean by this frequency domain analysis. We have the concepts of the wave lengths and the amplitude and the periodicities; we express the X t time series as a combination of cosine waves and sin waves. And there is a random component associated with this, n is a sample size here, I will discuss this in detail in the next lecture. But just to give you a flavor of what we do in the frequency analysis X t or time series we express this as consisting of cosine and sin waves with different harmonics f k is the k-th harmonic of the natural frequency.

And then, we do the analysis on X t by determining alpha naught and alpha k and beta k, and capture how much of variance that is present in a given frequency interval. So, essentially the idea is that how much of the variance in the data can be explained by different bands of frequencies. So, this is the basic principle of the frequency analysis; we will deal with the frequency analysis in detail in the next lecture.

So, we will just recapitulate what we covered in this particular lecture, in the lecture number twelve. We started with the stationary first order Markov model for generation of the data. This is specifically used for annual flows; generation of annual flows the parameters and the moments that it preserves are the mean standard deviation and the lag one correlation. Now, these are assumed to be stationary.

That is the model is stationary with respect to mean standard deviation and lag one correlation. Then, we relax this and build a non stationary first order Markov model which is typically used for generation of monthly flows and often it is also used for smaller time periods like 10 days time periods and larger time periods for seasonal flows. We saw three examples of using the non stationary model, first with the monthly data as given and then we convert that that into a log transformed data, the same non stationary Markov model can also be used for log transformed data.

If the log transformed data can be approximated using a normal distribution. Then we saw an example of an Indian case study, where we considered the flows into Sardar Sarovar Reservoir and the time duration that we were interested in that particular case was 10 days duration, a year was divided into 36 time periods and we generated several sequences. We generated sequences of large lengths of data and then compare the historical mean, standard deviation and lag one correlation with the observed data. Then, we consider several issues that we need to be alert too when you are choosing a stochastic model especially those dealing with peak flows, the time during which the peak flow occurs and so on.

Towards the end of the lecture, I just introduced the frequency domain analysis that means from the time domain, we convert the data into frequency domain and then start looking at the data in the frequency domain essentially to identify the periodicities inherent in the data. So, the purpose of all these methodologies that we are introducing in this course is to learn from the data - the data that is actually observed at a particular location is telling us some story about what has happened, we want to extract this information, so that we can model the data for purposes of several applications. So, we continue this discussion in the next lecture, in which I will introduce in detail the spectral analysis or the frequency domain analysis.

Thank you for your attention.