

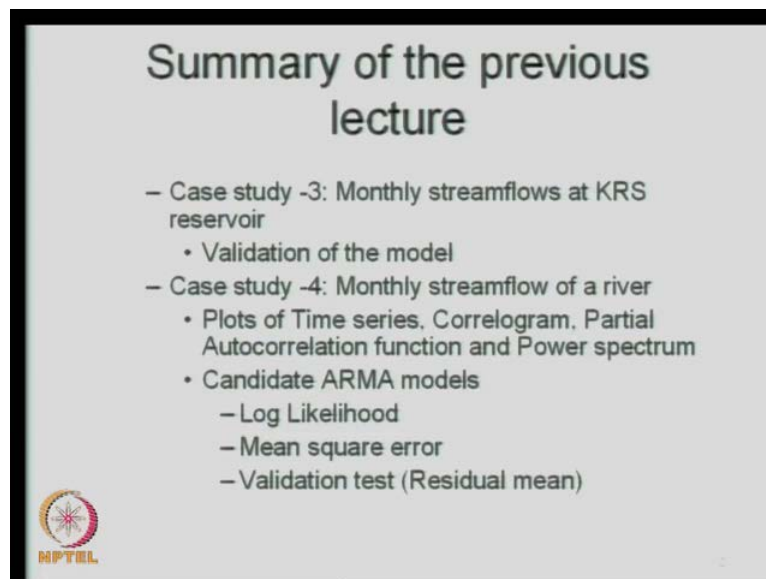
Stochastic Hydrology
Prof. P. P. Mujumdar
Department of Civil Engineering
Indian Institute of Science, Bangalore

Lecture No. # 21

Case Studies – IV

Good morning, and welcome to this the lecture number twenty one of the course stochastic hydrology. In the last lecture we dealt with the case study of the kaveri river flows at the KRS reservoir, we considered the monthly stream flows of the kaveri river. And then we plotted the time series, the correlogram, the spectral density.

(Refer Slide Time: 00:37)

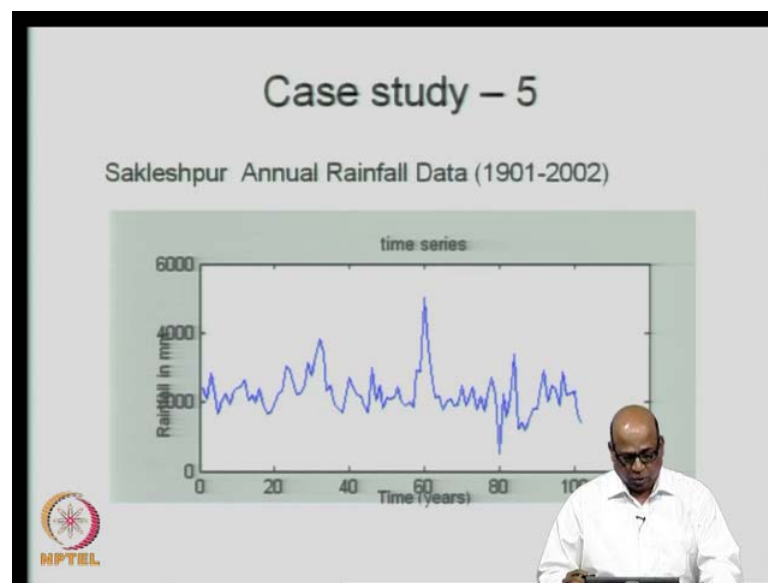


And then we went on to identify the model and we also, looked at the validation of the model both for synthetic generation as well as, for one time step ahead forecasting. Then we also considered the case study, of a stream flow in US river, American river where the metrologies are entirely different. And therefore, the contributions to the stream flow will be much different compared to the monsoon stream flows that we have in our part of the country. For the case study four again, we plotted the time series we plotted the correlogram, partial autocorrelation function and the power spectrum. Then we

formulated several candidate models ARMA of the ARMA type. For each of the model, we computed the log likelihood function, the mean square error.

And thereby, identifying a model for synthetic generation of the stream flows as well as for, one time step ahead forecasting using the mean square error criteria. Then we formulated the residual time series, residual series and then on the residual series we did the validation test. Towards the end of the last lecture, I just introduced the case study of the Sakleshpur rainfall. The Sakleshpur region comes on the Western Ghats and therefore, it has very high intensity of rainfall compared to the rest of the country in the rest of the country. So, we will continue that case study today and see how the power spectrum appears like and we also like go on to build an ARMA type of model for the Sakleshpur rainfall

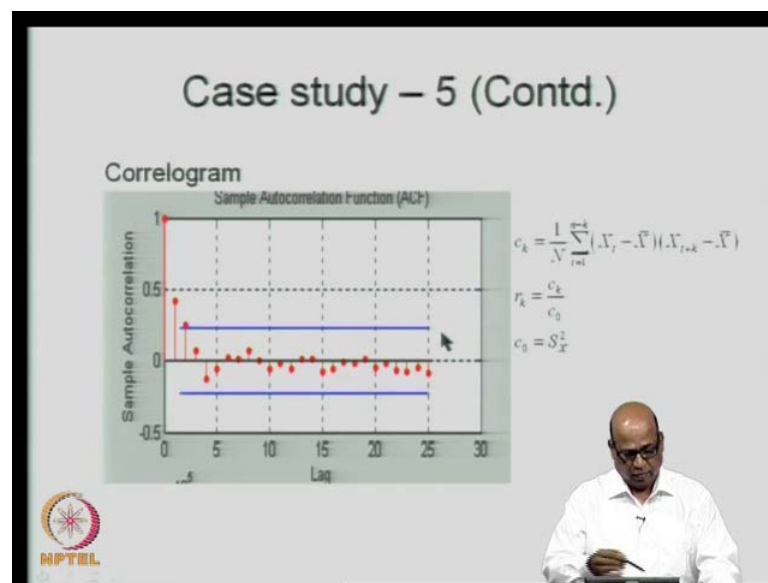
(Refer Slide Time: 02:34)



As, I mentioned earlier, you know we are considering the annual rainfall data here, but, if you have let us say, monthly data or daily rainfall data. Typically the rainfall as the duration becomes shorter and shorter. It becomes more difficult to analyze the rainfall with linear models, as we are doing now. So, typically the rainfall for shorter durations specifically, daily time steps or even weekly time steps etcetera they are not amenable to analysis by linear models. But because, we are considering annual rainfall data let us see, how the process itself behaves?

Now, this is the time series, compare this with the monthly time series of knavery river flows. For example, right at the time series plot you would know that, there are inherent period as it is present in the data. We have about hundred and two years of data for the Sakleshpur annual rainfall. And this 100 and 2 years data does not show at least, by examining the time series plot it does not indicate that there are any significant periodicities present in the data. But Let us see, what happens if we plot the correlogram?

(Refer Slide Time: 03:57)

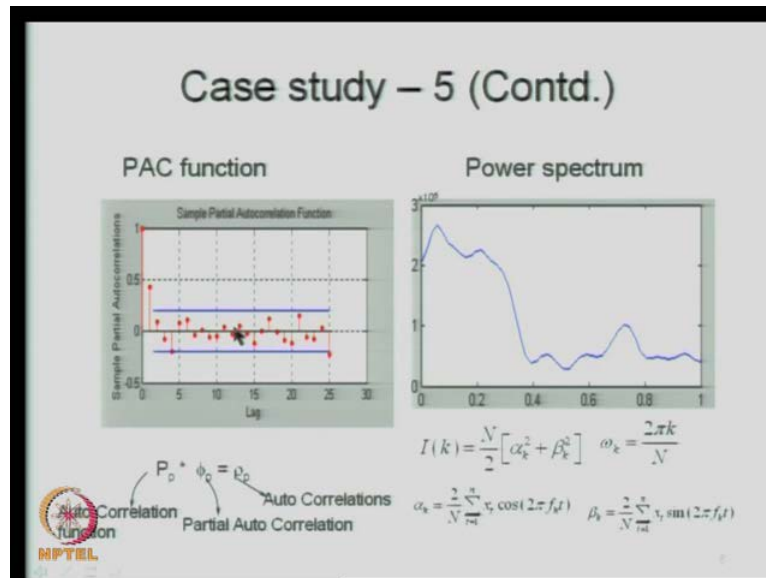


The correlogram for this again indicates that there are no periodicities. In fact, there are these are the significant bands, 95 percent significant bands. Most of the correlations up to about, like of 25 we have considered here, they are all insignificant only, the first two and perhaps the third one appear to be significant. In fact, this is lag zero. So, you leave about lag 1 and 2 seem to be significant in this whereas, all the other correlations are insignificant. Because, I have discussed correlogram, I think about 4, 5 lectures ago, let us how we recall? How we formulate this?

This is c_k which is the co variance at lag k , you are considering the co variance between x_t and x_{t+k} . So, $x_t - \bar{x}$ into $x_{t+k} - \bar{x}$ by N . And then r_k which is a estimate for lag k correlation is implicated as c_k by c_0 , where c_0 is the variance s_x^2 of the process, that you are considering. So, this is how you get r_k . So, this is a plot between r_k and k and typically, we go up to about 25 percent of the data. So, $0.25 N$ have gone for the correlation. So, the correlation plot or the correlogram

indicates that most of the correlations are insignificant you may only have 1 and 2 as significant and they are also decaying as you go with the lag. The correlations are decaying with the lag.

(Refer Slide Time: 05:45)



Let us, look at the partial auto correlation again as you all can see here, there is only 1 partial autocorrelation that is significant. That is at lag 1 these are 95 percent significance bands similar, to the correlogram how do we formulate this? This is by taking 1.96 by N on either side of 0. Or you can use it as a approximation 2 by root N that is 1.96 by root N or simply 2 by root N you can take and N in this case is 100 and 2. 100 and 2 years of data is available therefore, N becomes N that is how, you formulate the significance bands. The same significance bands you also use it for partial autocorrelations because, the number of data is still the same as you can see the partial autocorrelation at lag 1 only is significant.

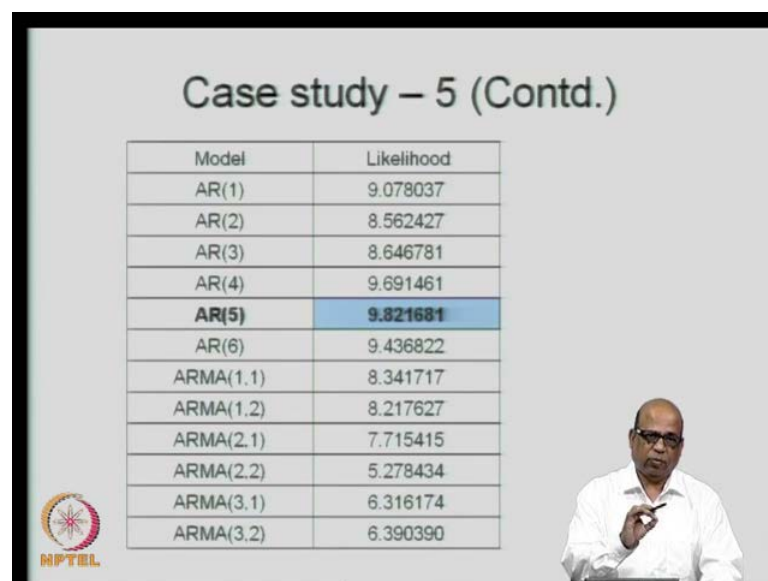
And at all other lags it is all partial autocorrelation is insignificant. How do we get the partial autocorrelation? Recall again that we formulate the Yule walker equations. Yule walker equations will give you from this, p into phi p is equal to rho p is the Yule walker equation. Where p is the autocorrelation function and phi p is the partial autocorrelation and rho p is the autocorrelations. When you formulate the Yule walker equation of order p, you will get the partial autocorrelation for at that order phi p at that particular lag. So,

corresponding to each of these lags you formulate the corresponding Yule walker equation and get the associated partial autocorrelations.

Then we also, formulate the power spectrum. The power spectrum again, you compare this with the power spectrum that you obtain for monthly stream flows. This is annual rainfall that we are considering and in a Western Ghats region. So, this is how the power spectrum appears again, recall how we formulate the power spectrum. I k this is N is 100 and 2 in this case alpha k you get and beta k you get from these expressions where x t is your time series, this is time series f k is 2 pi by n. So, you get corresponding to a particular k. You get f k and t is the time and x t again is the time series and N is 100 and 2.

And therefore, you get alpha k and beta k and you get omega k which is 2 pi k by N and I am sorry f k is k by N. If you recall from your spectral analysis f k is k by n. So, associated with any given k you can get f k and t is simply time over which you are summing it up and then x t is the original time series. So, you can get alpha k and beta k once you get alpha k and beta k you get I k. And omega k is given by 2 pi k by N corresponding to each of this k and this is the power spectrum which you get on this then.

(Refer Slide Time: 09:03)



Case study – 5 (Contd.)

Model	Likelihood
AR(1)	9.078037
AR(2)	8.562427
AR(3)	8.646781
AR(4)	9.691461
AR(5)	9.821681
AR(6)	9.436822
ARMA(1,1)	8.341717
ARMA(1,2)	8.217627
ARMA(2,1)	7.715415
ARMA(2,2)	5.278434
ARMA(3,1)	6.316174
ARMA(3,2)	6.390390

We will see again, like we did for the earlier case studies we formulate a number of candidate models. Now, the correlogram indicates that it is decaying either in a

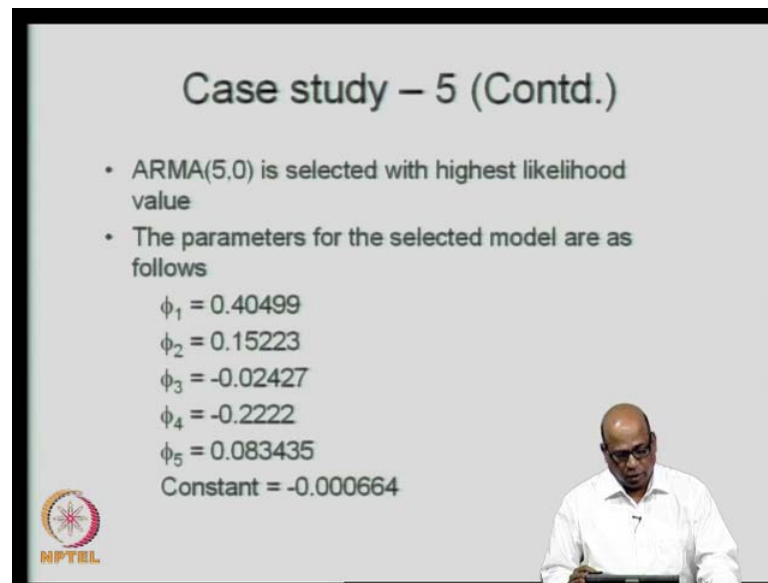
sinusoidal way or in an exponential manner the correlogram may be decaying. And there may be one or two significant partial autocorrelations. Based on this we suspect that this follows an AR model, AR type of model. That is autoregressive type of model what we then do is. In fact, we could have used straight away the AR 2 model based on the fact that, there are two partial autocorrelations which are significant and then the correlogram is decaying slowly.

In fact, AR one in this particular case because we may be saying that there is only one partial autocorrelation which is significant as you recall. There is just one partial autocorrelation which has also just on the border line if you look at the partial autocorrelation you may have this also as a significant and this one is also significant. So, you may have one or two partial autocorrelations significant you have the option of simply choosing based on correlogram as well as partial autocorrelation. But, typically what we do as, I mentioned in the last lecture is that we formulate a number of candidate models and then look at the likelihood values.

So, the candidate models we consider in this case are AR 1 to AR 6 and then ARMA 1.1 to ARMA 3.2, 1.1, 1.2, 1.1 etcetera up to 3.2. Associated with each of these models first we compute the parameters for taking by taking first half of the data, in the 100 and 2 years data we take let us say about 50 years of data. And then compute the parameters that is the model calibration. So, calibration we do with the first half of the data. And the parameters are estimated based on the first half of the data then we apply the model to the second half of the data. And get the associated likelihood values remember the likelihood value will have a \bar{e} associated with it that is the residual mean and so on.

So, residual wave mean and variance you will get the residual, you will get by applying the model any particular model to the validation data part. In this case, the remaining 50 years or 52 years etcetera, depending on how you have considered for the calibration. We obtain the likelihood values and we see, that AR 5 model gives the maximum likelihood. And therefore, we choose AR 5 for long term synthetic generation of the data. So, this is for the simulation of data.

(Refer Slide Time: 12:13)



The slide displays the following information:

- ARMA(5,0) is selected with highest likelihood value
- The parameters for the selected model are as follows
 - $\phi_1 = 0.40499$
 - $\phi_2 = 0.15223$
 - $\phi_3 = -0.02427$
 - $\phi_4 = -0.2222$
 - $\phi_5 = 0.083435$
 - Constant = -0.000664

The slide also features the NPTEL logo in the bottom left corner and a small inset image of a man in a white shirt in the bottom right corner.

Then we also consider, fine before I come to the forecasting part. So, AR 50 or simply AR 5 that is ARMA 50 AR 5 this model gives you the maximum likelihood value. The associated parameters for these are phi 1 to phi 5 and the constant. So, these are the associated parameters recall that we estimate the parameters by the marquardt's algorithm or by the r max function in the mat lab. So, when we do that and when we were computing the likelihood values the parameters for all of these models would have been estimated already. And with those parameter models parameters in place we would have simulated the data and for the remaining part of the data that is N by 2 values.

And then obtain the associated likelihood values. Now, to summarize the ARMA 50 model has phi is equal to 0.40499 and so on up to phi 5, and then the constant turns out to be minus 0.30664. Now, we have identified this particular model to be suitable for long term synthetic generation of the data. And we also, have the residual series with us. Now, we do the validation test on the model what are the three validation test if you recall? First of all, the residual should have a 0 mean. So, the series of residuals that you get you must first examine for the significance of the mean.

And must satisfy yourself, that you can approximate the mean of the residual to be equal to zero. Then we look for the periodicities whether there are any significant periodicities present in the residual series. Then we also, test whether the residual series that you

obtain by applying a particular model is. In fact, a white noise in the sense that the residuals are all uncorrelated. So, these are the three tests that we do.

(Refer Slide Time: 14:35)

Case study – 5 (Contd.)

- Significance of residual mean

Model	$\eta(e)$	$t_{0.95}(N)$
ARMA(5,0)	0.000005	1.6601

NPTEL

So, first for the residual mean recall from the last lecture how we obtain this eta e? And this is again for the residual series. And for the 95 percent confidence level the t value t distribution value N is in this particular case is hundred and two. So, you get for 100 and 2 values this is N by 2 actually. So, we go with 52 values, because we have considered the first 50 values as for calibration. So, the remaining 52 values we are taking it for validation. So, you get t 0.95 corresponding to this as 1.6601 and therefore, the model passes the test because the statistic that you have computed is less than the critical value of t and therefore, the model passes the test.



What do you mean by the model passes the test? It means that, the residual series has a zero mean. Or you can be confident, that the residual series insignificant, it has a insignificant mean then.

(Refer Slide Time: 15:45)

Case study – 5 (Contd.)

Significance of periodicities:

Periodicity	η	$F_{0.95}(2, N-2)$
1 st	0.000	3.085
2 nd	0.00432	3.085
3 rd	0.0168	3.085
4 th	0.0698	3.085
5 th	0.000006	3.085
6 th	0.117	3.085





We look at the periodicity again, the first test of the periodicity this is not the cumulative period gram test. This is the first test that I introduced in the last lecture, where you look at the statistic eta. And then compare it with f statistic, the critical value of f statistic with two degrees of freedom N minus two. Again, N is in this particular case 52. So, you take 52 minus 2 and then get the associated f value that is 3.085. And the eta values that you get, corresponding to each of these periodicities, first periodicity, second periodicity etcetera. We will be all less than the corresponding f and therefore, the residual series that you obtain from the application of the ARMA 50 model they are all insignificant. It has insignificant periodicities and therefore, it passes the test.

(Refer Slide Time: 16:52)

Case study – 5 (Contd.)

- Whittle's white noise test:

Model	η	$F_{0.95}(n1, N-n1)$
ARMA(5,0)	0.163	1.783




Similarly, we look at the white noise test and if, you recall it is the Whittles white noise test in which we formulate the eta again. And then compare it with the f distribution with arguments this is a degree of freedom $N - 1$ which is a k_{max} . k_{max} in this particular case we have taken it as 25 the lag maximum lag that we have considered. So, this is 25 and $N - 1$. N is 52 and $N - 1$ is 25. So, you get $k_{f, 0.95}(n1, N - n1)$ as 1.783 and therefore, this eta being less than the critical value of f it passes the test. And we conclude that, the residual series that you obtain, by applying the ARMA 50 model on the remaining half of the data passes the test of white noise. It indicating that the residuals are all uncorrelated.

(Refer Slide Time: 17:58)

Case study – 5 (Contd.)

Model	MSE
AR(1)	1.180837
AR(2)	1.169667
AR(3)	1.182210
AR(4)	1.168724
AR(5)	1.254929
AR(6)	1.289385
ARMA(1,1)	1.171668
ARMA(1,2)	1.156298
ARMA(2,1)	1.183397
ARMA(2,2)	1.256068
ARMA(3,1)	1.195626
ARMA(3,2)	27.466087

Handwritten notes:
 $e_t = x_t - \hat{x}_t$
 $MSE = \frac{\sum_{t=1}^N e_t^2}{N}$
Validation Period



Then we look at remember what we did so far, for this case study is for long term synthetic generation of the data. Then we are interested in short term one time step ahead forecasting, we use the minimum mean square error criteria. So, we obtain corresponding to each of these models, we obtain the mean square error. And then compare the mean square errors and pick that particular model which has the maximum, which has the minimum mean square error. In obtaining these mean square errors again we use the same parameter value that we had obtained earlier. In this candidate models we would have obtained the parameters. The same parameters we obtain and then we apply these models, for the remaining half of the data and compute the mean square errors.

If, you recall the error of forecasting we are applying the model now, one time step ahead. Let us say, this being a yearly data let us say, we stand in a particular year 1985. Let us say, we stand at the end of that year and then forecast for the next year 1986 using these models any of these models. Because, we already have the data for that particular year the forecasted value minus the actual observed value gives the error. Or the actual value minus the forecasted value does not matter because, we are taking the square of that that gives the error e_t and that e_t we square it and get them. So, here we get e_t as x_t minus \hat{x}_t this is the forecasted value this is the actual data and this we are doing it for the validation time.

So, e_t you get and then e_t^2 you take divided by N is equal to 1 to n . So, this is how, you get the mean square error that is the mean square error that you are obtaining this is for the validation period (O). And then when we apply these models this \hat{x}_t or \hat{x}_t cap is obtained by applying the a particular model forecast arising out of that particular model is denoted as \hat{x}_t cap there. And we choose here, the ARMA 1.2 model as the best model, arising out of the mean square error criteria. And this is what we use it for forecasting one time step ahead forecasting. Let us see, what are the parameters that we obtain for this.

(Refer Slide Time: 20:58)

Case study – 5 (Contd.)

- ARMA(1, 2) is selected with least MSE value for one step forecasting
- The parameters for the selected model are as follows

$\phi_1 = 0.35271$
 $\theta_1 = 0.017124$
 $\theta_2 = -0.216745$
 Constant = -0.009267

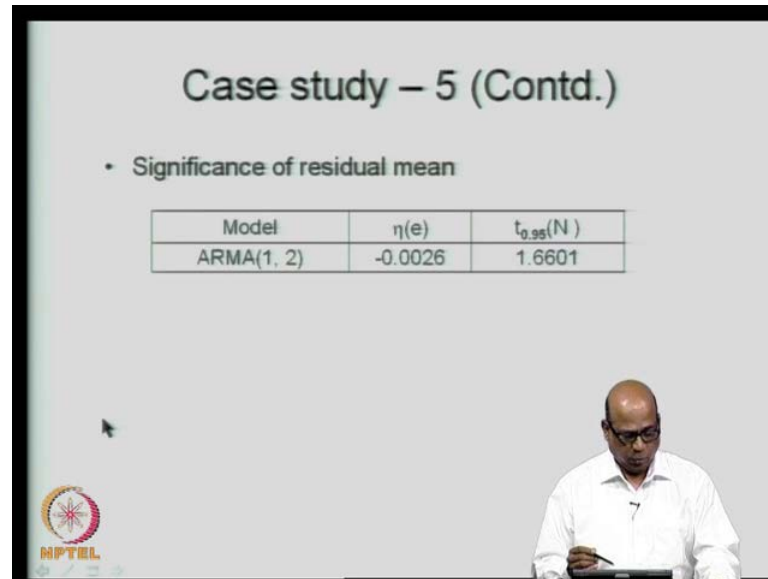
$$X_t = \phi_1 X_{t-1} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + e_t$$

So, ARMA 1.2 we selected and these are the parameters ϕ_1 , θ_1 , θ_2 how did we write the ARMA 1.2 model? So, the ARMA 1.2 model in this case we write it as x_t you have 1 AR parameter. So, I will write it as $\phi_1 x_{t-1}$ and you have 2 MA parameters. So, $\theta_1 e_{t-1} + \theta_2 e_{t-2} + e_t$ this is our ARMA 1.2 model. So, you get ϕ_1 , θ_1 , and θ_2 . So, these are the values $\phi_1 = 0.35271$, $\theta_1 = 0.017124$ and $\theta_2 = -0.216745$ and constant is -0.009267 . I must, alert that we need to consider several digits of ϕ_1 , θ_1 , θ_2 etcetera because, the x_t will be quite sensitive to these parameters.

And therefore, you consider several digits like 3 or 4 or 5 digits you consider for each of these parameters. Now, we apply this model now ϕ_1 is known, θ_1 is known, θ_2 is known and e_{t-1} , e_{t-2} etcetera as I have explained on the application of

the ARMA models. You use these as errors of the previous forecast for example; e_t minus 1 is the error of the forecast for t minus 1, e_t minus 2 is the error of forecast for t minus 2. So, like this you do, then when you are taking the forecast what you do the expected value? You take the expected value and therefore, this term vanishes because, it has a zero mean.

(Refer Slide Time: 23:00)



Case study – 5 (Contd.)

- Significance of residual mean

Model	$\eta(e)$	$t_{0.95}(N)$
ARMA(1, 2)	-0.0026	1.6601



So, apply this for forecasting and formulate the error series e_t and then we do the tests on the error series or the residual series. So, the first two test that we can conduct on the residual series is a significance of the residual mean whether, the we can approximate the mean to be zero. And we compute the statistic, compare it with the associated t value much the same way as we did for the calibration part for the I am sorry not for the calibration part. It is for a long term synthetic data where, we choose the ARMA 50 model. Similar test we conduct of the residual mean for the forecasting model also, then we also look at whether, the residual mean has any significant periodicities present in that.

(Refer Slide Time: 23:52)

Case study – 5 (Contd.)

Significance of periodicities:

Periodicity	η	$F_{0.95}(2, N-2)$
1 st	0.000	3.085
2 nd	0.0006	3.085
3 rd	0.0493	3.085
4 th	0.0687	3.085
5 th	0.0003	3.085
6 th	0.0719	3.085



So, we do the significance of periodicities we examine several periodicities typically these periodicities that we examine are obtained either from your time series plot directly or from the spectra density. In the spectral density if you suspect that there are periodicities associated with let us say, 10 years 15 years etcetera, based on the visual observation you consider all those periodicities. So, typically we do for 5, 6 periodicities which are emanating from the spectral density and then obtain for the residual series obtain the eta values compare them with the f statistic in this particular case 2 and N minus 2.

So, N is again 50 years of data and then obtain compare the eta value with the f distribution to ensure that all the eta values that you have are less than the critical f value and therefore, it passes the test for periodicity.

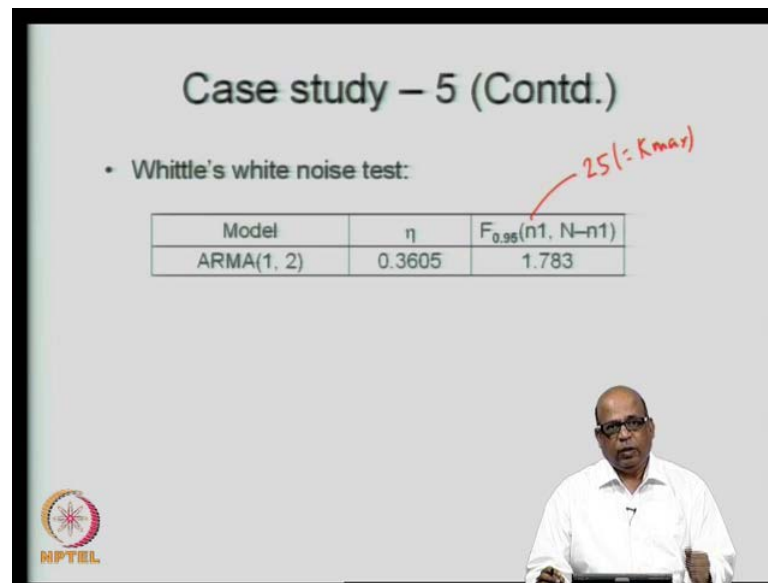
(Refer Slide Time: 25:09)

Case study – 5 (Contd.)

- Whittle's white noise test:

Model	η	$F_{0.95}(n1, N-n1)$
ARMA(1, 2)	0.3605	1.783

25 (= k_max)

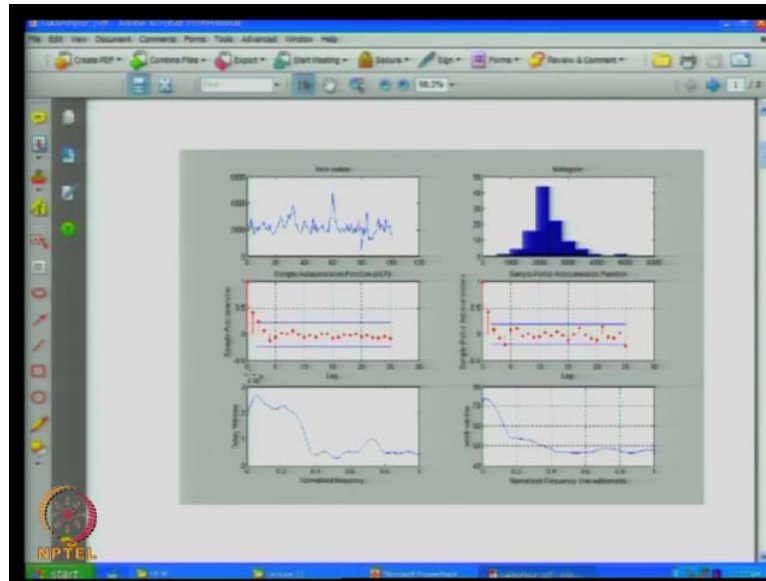


Then the last test we do is for correlation that is we have to make sure that the residual series that you obtain is. In fact, a white noise or the residues are all uncorrelated. So, this we do by Whittles white noise test and this is a statistic we obtain 0.3605 compare this for remember for the Whittles test you take $N-1$ and N minus $N-1$. Now, $N-1$ is the maximum lag in this particular case, it is as I said, we would have taken 0.25 of N about 25 or something. So, $N-1$ in this case may be 25 this also equal to k_{max} that is the maximum lag.

So, you get the critical value of f as 1.783 and whereas, you're your η the description of this η I have covered it in the last lecture please refer to the Whittles white noise test and the associated η value is 0.3605. And therefore, this passes the best because η is less than the associated critical value of f . And we conclude that the model that we have chosen namely ARMA 12 for one time step ahead forecasting passes all the tests on the residual series. And we choose, this particular model for one time step ahead forecast what I will do now, is for this particular case study 5, I will give you all the details of including let us say, including how we formulate the models.

And the intermediate results etcetera let me just go through the case study in some detail. This is typically we give this kind of projects as part of a course. So, when I teach stochastic hydrology for the masters degree at Indian institute of science these projects are done by the course students. So, we give 10 percent or 15 percent weight age for this.

(Refer Slide Time: 27:39)



So, the Sakleshpur rainfall data is given I will also, supply with the data with this and then first you look at how the data appears by itself. Let us say, you are looking at the time series all of this I have explained. You also, formulate the histogram this is the original time series, you form the histogram how do I formulate the histogram? You divide the data into several class intervals. And look at the frequency how many number of values have appeared in let us say, between 1000 to 2000 and so on. So, like this you formulate the histogram.

The histogram along with the time series gives you an indication of how the values are distributed then you form the correlogram this is a correlogram associated with this. And then you also look at the partial autocorrelation then you look at the spectral density. Now ,in this particular case the spectral density has been, computed both with the tukey window as well as the welsh window.

(Refer Slide Time: 28:58)

Sno.	Model	MSE	Max-Likelihood
1	AR(1)	1.180837	9.078037
2	AR(2)	1.169667	8.562427
3	AR(3)	1.182210	8.646781
4	AR(4)	1.168724	9.691461
5	AR(5)	1.254929	9.821681
6	AR(6)	1.289385	9.436822
7	ARMA(1,1)	1.171668	8.341717
8	ARMA(1,2)	1.156298	8.217627
9	ARMA(2,1)	1.183397	7.715415
10	ARMA(2,2)	1.256068	5.278434
11	ARMA(3,1)	1.195626	6.316174
12	ARMA(3,2)	27.466087	6.390390

MODEL VALIDATION RESULTS

MODEL AR(5)

The tukey window is what I have explained in the class. So, look at these entire all of these figures. These are for the original time series then we compute the MSC as well as the maximum likelihood values. So, this is why I am giving you all of this again repeatedly is, that I will give you also the procedure by which you obtain these or residual series as well as for forecasting model and for the synthetic generation model. So, you get the MSC as well as, the maximum likelihood values to consider the 12 candidate models then, we do the model validation. First you would have got the parameters for AR 5 these are the model parameters 1 2 3 4 5. 5 model parameters for AR and then there is a constant this is a constant there.

Then we do the significance test first the test for residual series with to examine whether, this has a zero mean you get the eta value and the associated statistic. And then look at the result whether this eta value is less than the associated statistic. Then you look for statistical test for significance of periodicity, corresponding to each of these periodicities. You look at the eta value and the associated $f_{0.95}$ and make sure that all is passed the test. Then we look at Whittles white noise test all of this I have discussed. So, let me go a bit fast we have also, introduced the turning point test which I did not discuss in this particular course.

But you can always go with the Whittles white noise test for the randomness then we look at the residuals. So, once the residuals have passed the entire test we also, look at

the residuals properties themselves. The histogram of the residuals the sample autocorrelation function or the correlogram as well as the sample partial autocorrelation these are for the residual. Now, there is a concept called normal probability plot, which I had not discussed. So, far in the class may be towards the end of the course I will introduce this. This is to examine whether particular data series. In fact, follows normal distribution or not.

So, for the time are you do not worry about this look at these 3 figures. So, this is on the residual series then again you look at the ARMA 12 model which we selected for one time step ahead forecasting do all the tests again, do the test on residual series to examine that it has a 0 mean it has insignificance of periodicities. And then it also passes the test for randomness or the white noise test and then we also look at the plots of the residuals themselves now these plots of the residuals as well as the original time series give us an important insights into how the particular model is behaving now what we do is

Let's say we have chosen the AR 5 models as the model to be used on this particular case study for long term simulation of data. So, we apply that model for let us say about 100 years 100 and 50 years. We simulate the data and then compare the various statistics of the simulated data with those statistics of the original data. Essentially what we do now, we have chosen AR 5 model for synthetic generation of the data. We have estimated all the parameters, we have convinced ourselves that this model passes all the tests. And therefore, this model is ready for application let us apply this model now. So, we apply this model to simulate a large number of data belonging to that particular time series

So, and then we compare the statistics. So, remember this is for the original time series and then we have simulated it is not for validation period. We have done all the validation test kept the model ready now we simulate for next hundred years 1 sequence 2 sequence etcetera like this next 100 years we simulate the data and then compare the data the mean of the original simulated series with the original series. So, this compares like this similarly, the standard deviation compares 488 590 then there is a skewness minus 0.01867 and this is 1.167 kurtosis 2.90 7.96 it may appear that the model is not reproducing well the skewness and kurtosis, but, it was not meant to.

In fact, you see there was no feature in the model that that would reproduce this statistic, but; however, for comparison purpose we have still used these because in certain situations you would like the particular model to reproduce certain statistic for example, skewness. If the skewness in the original data was significant you would like to preserve that kind of a sequence skewness in the final simulated data also in which case it is better to compare them. So, while you have not built the model to preserve the kurtosis and the skewness specifically, but, the model that you are actually applying for long term synthetic data you would be curious to see whether the skewness is preserved at all.

And in this particular case skewness as well as the kurtosis are not well simulated whereas, the mean and standard deviation we can take them to be acceptable. So, we now compare how the original time series and the synthetic data appear to be. So, this is for the original time series the left side what you are seeing is for the original time series these are all for the original time series and this is for the simulated data. You can see some similarities not. So, much in the time series itself, but, if you look at the histogram **histogram** appears to be similar in some sense here in terms of the distribution of the data remember here the scales are different 500 to 3 7 5 100 100 whereas, here it is 0 to 6000.

So, that is why it appears to be slightly wider in this case. Then you also look at the partial autocorrelations the partial autocorrelations seem to be I am **sorry** this is a correlogram **correlogram** seems to be reproducing it well. In fact, the time series model this is AR 5 models therefore, the information on the correlation will be well preserved in the original data in the simulated data with respect to the original data. Now, in this file here PDF file I have given you the complete data this is the station is Sakleshpur state is Karnataka basin is Hemavathy the source is earlier it used to be called as a drought monitoring cell now it is called as disaster management cell or some similar name it is located Bangalore.

So, we have obtained the data from this source and this **this** data I am giving here you can use this as an exercise. So, take this data do the complete time series analysis you have the data from nineteen hundred and one to two thousand and two here. So, you use this data and do the time series analysis that I have discussed in this particular case study and satisfy yourself that you get all of these values correctly.

Alright now we will try to summarize what we did in the case studies let me give the summary of the results that we obtain for various case studies. So, that you get a physical feel remember I have discussed the case studies only for rainfall and stream flow at says different time steps.

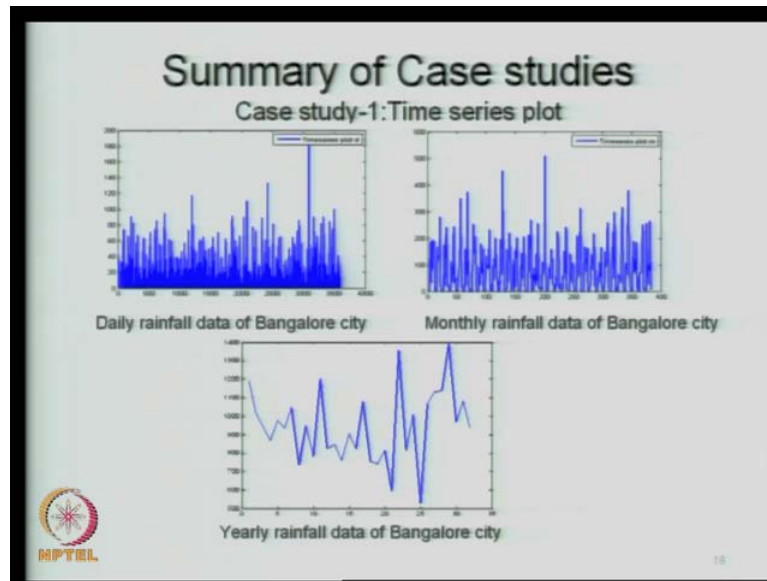
But you should be able to do the same analysis on let us say evapotranspiration if you have large number of data corresponding to evapotranspiration you can do the same type of analysis several cases do exist in literature where the time series analysis has been applied to water quality parameter water quality indicators for example, dissolved oxygen at a particular location.

If you have a series of dissolved oxygen values at a particular location and **and** the stream has a stabilized you in the sense that in terms of the effluents that it is receiving and in terms of the hydrology and etcetera it is stabilized and you are only looking at the particular water quality indicator time series then you choose that data and then apply the data to carry out the analysis that we have discussed.

Similarly, there may be other variables for example, ground water fluctuations if you have ground water levels at a particular location for a significant length of time then you use that as a time series apply the techniques that we have just discussed and then build models for forecasting how the ground water levels at a particular at that particular location are likely to be held.

And similarly, you can build models for synthetic generation of the ground water level. So, many of the hydrologic variables that we will be interested in **in** making decisions or in making long term plans can be analyzed using the time series techniques that we have discussed. So, before we close the topic on the time series analysis let us summarize the case studies themselves. So, first we consider the Bangalore city daily rainfall data. So, this is how the daily rainfall data appears let me come back to that.

(Refer Slide Time: 40:23)



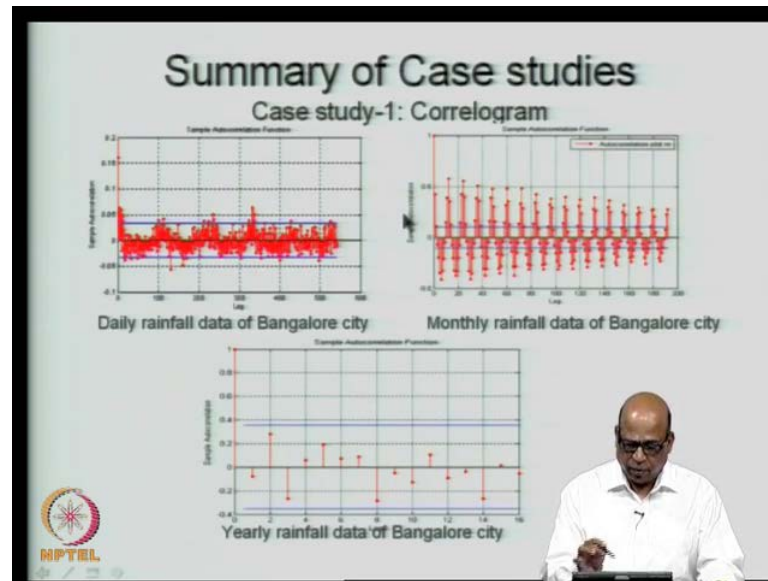
So, we started with daily rainfall then we went on to aggregate the daily rainfall to get the monthly time series then aggregated the monthly rainfall to get the annual rainfall time series.

So, this is how the daily rainfall data appears there are a large number of values and you can see that you may not be able to see any significant periodicities there is no periodicity when you are considering the daily rainfall data whereas, if you look at the monthly rainfall data there appears to be some smoothing and therefore, you may suspect that there may be some periodicities present here.

Again if you aggregate the monthly data and then put it into an annual time series again the periodicities have gone. So, both in the daily rainfall as well as in the annual data you may not see any periodicities then we formulated the correlograms corresponding to each of them let us the correlograms in the Bangalore city rainfall this was for the daily rainfall this is for the monthly rainfall and this is for the annual rainfall.

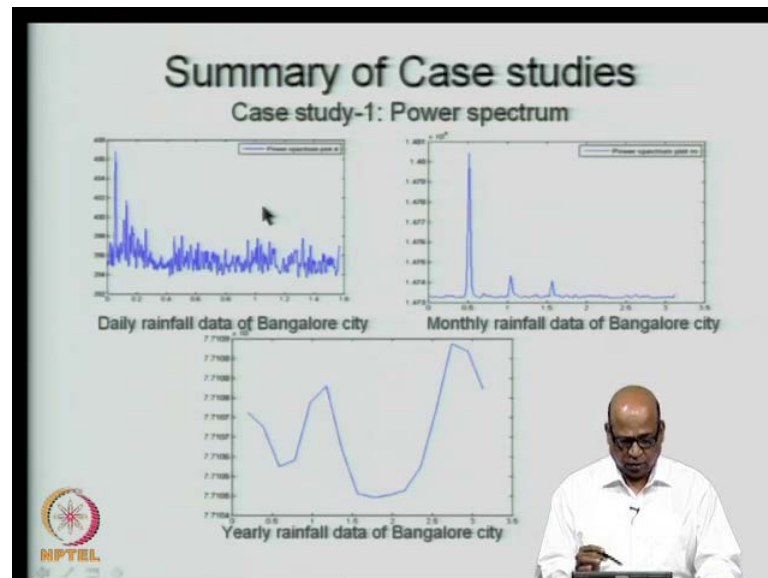
As was suspected in your time series data we thought that the daily rainfall may the monthly rainfall I am **sorry** may have some periodicities present in the data.

(Refer Slide Time: 41:36)



These are confirmed by the correlogram the correlogram indicates that there are periodicities here the correlogram is oscillating in a sinusoidal way. So, there are periodicities present here whereas, most of the correlations here in the daily case are all insignificant similarly, in the annual case they are all insignificant.

(Refer Slide Time: 42:05)

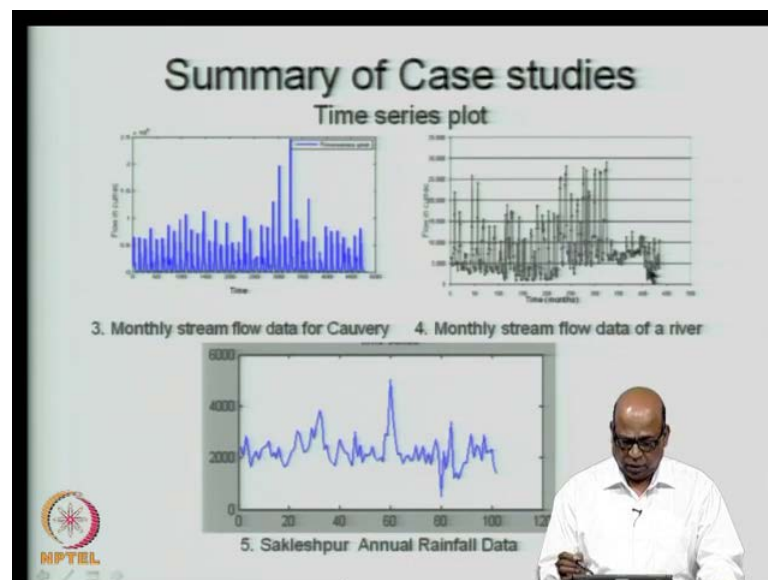


We went on to plot the spectral densities for each of these cases the spectral densities appear like this indicating that there is no there is no single frequency at which a large

contribution to variance takes place. So, this is more or less like a white noise whereas, the monthly periodicity monthly data throws up some significant periodicities.

One corresponding to twelve months is much is quite a significantly different from those corresponding to six months and four months whereas, again the yearly data does not show any preference for any particular frequencies the high frequencies seem to be dominating here, but, there are no significant periodicities thrown up by the yearly data spectral analysis.

(Refer Slide Time: 43:04)

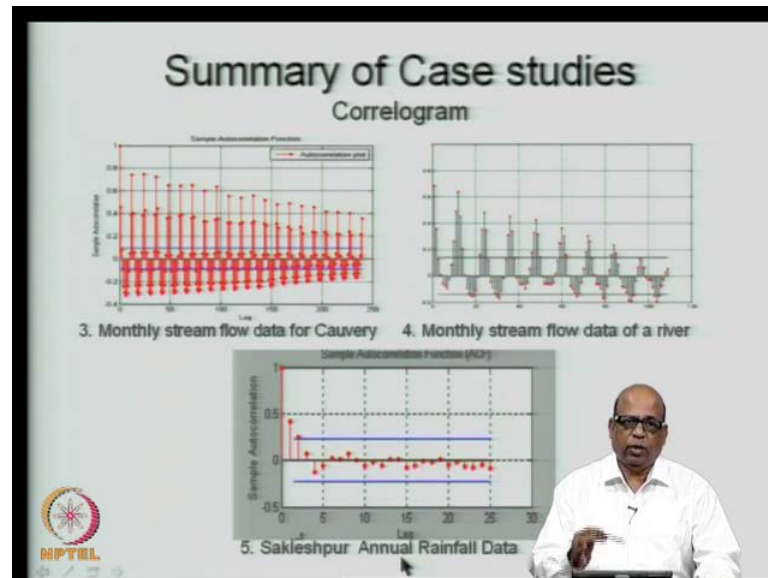


Then we consider the Kaveri river flow and the stream flow of a u s river here and also Sakleshpur annual rainfall data here. So, these were the time series plots this is just to show how the time series may appear differently for different types of variables and different types of locations for example, monthly times monthly stream flow time series of a h a monsoon climate typically appears like this.

Where you can see that there are some oscillations in the data year whereas, here it may show certain trend here that may be slight significant slight increasing trend that is seen here and this is the annual data annual data does not show up any periodicities or does it show any nor does it show any trend here

So, this is how the different time series appear on a visual inspection let us look at the correlograms also the monthly time series of the Kaveri river yields a periodic or the sinusoidal type of correlogram.

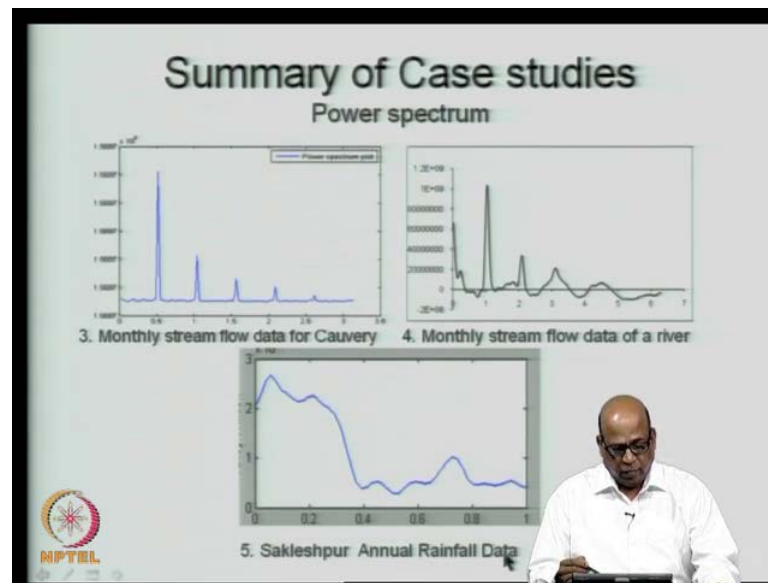
(Refer Slide Time: 44:35)



And the monthly stream flow data of the u s river this also shows periodicities, but, there are many of the negative periodicities which are negative correlations which are all insignificant and both of them show a slow decay.

This correlogram shows a slow decay here this also shows a slow decay in the correlograms the Sakleshpur annual data on the other hand shows there is only one significant correlation and all of them all of the remaining correlations are all insignificant remember lag 0 is always one the correlation at lag 0 is always one and therefore, you look at only lag one and beyond the all the correlations here are insignificant.

(Refer Slide Time: 45:31)



So, the rainfall typically behaves much different compared to the monthly stream flows this is what we observe from here then we look at the power spectrum power spectrum essentially confirms our **our** suspicion that there are periodicities present in the monthly stream flow data. So, you see that there is a periodicity associated with w_p of 0.52 or some such thing which corresponds to 12 months this is 6 months whereas, twelve periodicity is absent in the river that I have considered from u s. So, there is no there is no 12 month periodicity here.

Whereas there is a 6 months periodicity there is a 4 month periodicity now the 6 month periodicity appears to be quite significant for the monthly stream flow of a western river whereas, in the peninsular Indian river the 12 month periodicity almost always comes out to be significant whereas, the Sakleshpur annual rainfall data does not show one periodicity to be any different from any other periodicity although here it appears to be that the frequencies are dominating

(Refer Slide Time: 46:34)

Summary of Case studies
ARMA Models

3. Monthly stream flow data for Cauvery
ARMA(4, 0) – For data generation
ARMA(1, 0) – For one step forecasting
4. Monthly stream flow data of a river
ARMA (8, 0) – For both data generation & one step forecasting
5. Sakleshpur Annual Rainfall Data
ARMA(5, 0) – For data generation
ARMA(1, 2) – For one step forecasting

USGS site

MPTEL

And then we went on to build models the ARMA type of models for the monthly stream flow data for data generation we got AR 4 model and for one time step ahead forecasting we got the AR 1 model that ARMA 4 0 or ARMA 1 0 model then for the monthly stream flow data of a river whenever I say a river this is we have taken it from a from the web USGS site this is AUS river.

Ah. In fact, those of you who are willing to experiment on different data sets this is a good site from which you can download the data time series data and then do all your all the analysis that I have discussed in this course unfortunately such a data is not very freely available in India and therefore, you may not be able to do all this analysis that I have discussed satisfactorily with the Indian data unless you put extra effort and obtain data from several sources who are the custodians of the data in the country.

Now, for the monthly stream flow data of the u s river we got AR 8 0 remember here these type of models that we are getting themselves show the difference in the processes that have contributed to getting that particular time series. So, as I mentioned in the in the peninsular Indian region essentially the stream flows are generated by the monsoon rainfall.

And therefore, you get the significant periodicity every year you get some **some** kind of a regularity that monsoon months have higher flows compared to non monsoon months and. So, on whereas, if you look at time series of stream flows in let us say in European

region that may be much different from the peninsular Indian region which may be much different from a u s river and. So, on

So, it depends on how what kind of physical processes are. In fact, generating the particular time series that we have observed, and based on that you get different types of models coming up. Because the dependence of values observed at a particular time on those observed at some other time depends on the type of physical processes that are governing this governing that particular process.

So, you get a 8 0 model for both data generation and one time step ahead forecasting which may be quite surprising to **to** those who are working on especially peninsular Indian rivers where we typically get models of this type and especially for forecasting we often come with ARMA 1 1 or 1 0 model or maximum 2021 type of models or one time step ahead forecasting for whereas, per monthly stream flow data of the u s river you get a ARMA 8 0 model both for long terms simulation of the data as well as for one time step ahead forecasting.

Then for the Sakleshpur annual rainfall data remember rainfall data behaves much differently from the stream flow data stream flow is a much smoothed process we get a model of ARMA five 0 for data generation and ARMA one two for one time step ahead forecasting I had provided the data for Sakeleshpur annual rainfall.

And I encourage all of you to go through this data do all the analysis or that I have discussed in this particular course with that particular data if you do not have access to any other data use that particular data and then do all the analysis. So, with this now we conclude that the discussion on time series analysis.

Ah. So, this is a right time to just quickly summarize what we have covered in the time series analysis. So, we formed the series x_t and then we discussed about the stationarity of time series and we also introduced the concept of ergodicity of time series when the time average properties are same as the same ensemble average properties are the same as time average properties.

Then we say the processes ergodic and then we went on to build time series models we also discussed analysis in the frequency domain where we express the time series with

frequencies and then we did the spectral density we did the spectral analysis spectral analysis essentially shows up the periodicities present in the data

We discussed correlogram that is autocorrelations auto covariance function then we introduced the concept of partial autocorrelations and then went onto build arima type of models. So, first we have to convert the time series into a stationary time series. So, the models that I have discussed are all stationary time series models and the type of models the specific models that I have introduced are in addition they are all linear models.

Because how do I write the ARMA 1 0 models we write it as x_t is equal to ϕx_{t-1} plus e_t . So, they are all linear models there are no non-linear terms or we specifically introduce the differencing to make the non stationary time series to convert the non stationary time series into a stationary time series.

So, we looked at the first order differencing second order differencing and. So, on that is how do I write the first order differencing it is $x_t - x_{t-1}$ that is all. So, simply take the difference of a particular value with that of the previous value the second order differencing will be $x_{t-2} - x_{t-1}$.

So, that is how where x_{t-1} is the first order difference. So, this is how we do the differencing typically in most hydrologic time series if you do first order or second order differencing the series can be converted into a stationary time series; however, there may be periodicities present in the data.

In which case you may have to account for these periodicities in the time series models that you consider and that is where we look at both contiguous and noncontiguous type of ARMA models. So, we identify the AR components and MA components essentially by the plots of correlogram and partial auto correlation functions and also the spectral density.

So, there is a procedure of identification of the models then calibration in the calibration we discussed the estimation of the parameters although I did not cover the algorithm of the parameter estimation I have given you the `rmax` function the syntax which you can use on mat lab then estimate parameters for any of the ARMA type of model. So, first we do the differencing.

Arima remember arima is autoregressive integrated moving average models now that integrated I refers to the differencing. So, the order of differencing when I say arima p d q model we have AR terms of order p differencing of order d and m a parameters of order q that is how we write arima p d q models

So, first we do the differencing and then on the difference series you apply the ARMA models. So, r max function of the mat lab gives you the parameters of the ARMA type of models then we discussed how we obtain the likelihood values corresponding to each of the ARMA models and also the mean square error.

Which means that in the calibration period we take a part of the data typically the first half of the data and then estimate the parameters apply the particular models for the next half of the data validate the model by validation I mean you obtain the residual series and make sure that the residual series passes all the tests namely that there are the mean of the residual series is 0 there are no significant periodicities present in the residual series.

And also that the residual series is uncorrelated or it forms a white noise. So, these are the three tests that we do on the residual series. So, we now know for a given time series how to get a time series model or the ARMA type of model both for the synthetic generation of the data as well as for one time step ahead forecasting.

So, we will close this discussion today on time series analysis in the next lecture I will introduce the Markova chains which is a slightly different topic from what we have been doing. So, far. So, we will continue the discussion in the next lecture where I will start with discussion on Markova chains thank you for your attention.