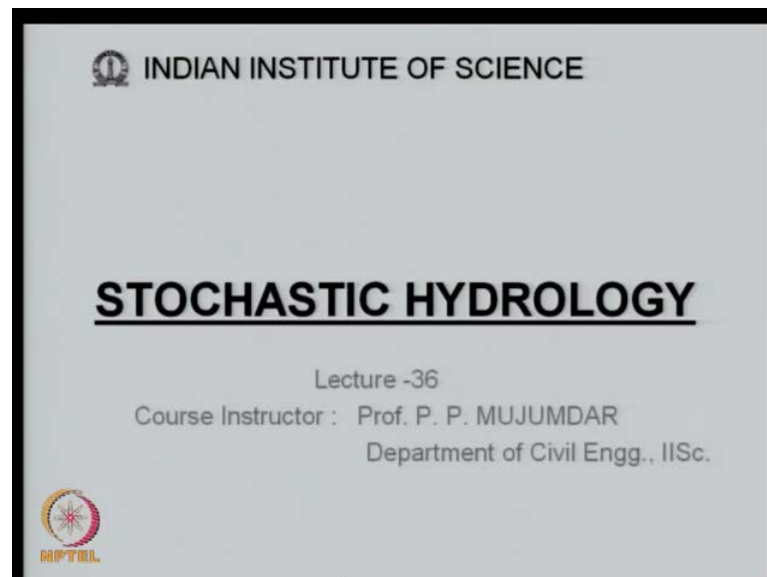


Stochastic Hydrology
Prof. P.P.Majumdar
Department of Civil Engineering
Indian Institute of Science, Bangalore

Module No. # 08
Lecture No. # 36
Data Consistency Checks – I

(Refer Slide Time: 00:24)



(Refer Slide Time: 00:28)

Summary of the previous lecture

- Multivariate stochastic models
 - Matalas model



$$X_{t+1} = AX_t + B\epsilon_{t+1}$$

where
 X_t and X_{t+1} are $p \times 1$ vectors representing standardized data corresponding to p sites at time steps t and $t+1$ resp
 ϵ_{t+1} is $N(0,1)$; $p \times 1$ vector with ϵ_{t+1} independent of X_t
 A and B are coefficient matrices of size $p \times p$. B is assumed to be lower triangular matrix

$$A = M_1 M_0^{-1}$$

$$BB^T = M_0 - M_1 M_0^{-1} M_1^T$$

M_0 is the cross-correlation (size $p \times p$) of lag zero
 M_1 is the cross-correlation (size $p \times p$) of lag one

Welcome to this the lecture number thirty six of the course stochastic hydrology. In the previous lecture we introduced the multivariate's stochastic models where essentially we are looking at data generation simultaneously at several locations; in a basin or in adjoining basins where the primary purpose is to preserve the cross correlation structure among the flows at various locations. And this we distinguished with respect to the single site models where the single sites models will essentially preserve the statistics of flows at that particular site alone in terms of the mean, the variance and the lag one correlation and so on.

In the multisite models along with preserving the means and standard deviations and lag one correlations at that particular site, they also preserve the multisite models. Also preserve the cross correlations with other sites in the same basin and we introduced the models of the type $x_t + 1 = a x_t + b \epsilon_{t+1}$ where a and b are the coefficient matrices. It is **in it is** through a and b that the cross correlations are accounted for. x_t is the data at time period t and ϵ_{t+1} is a random variate.

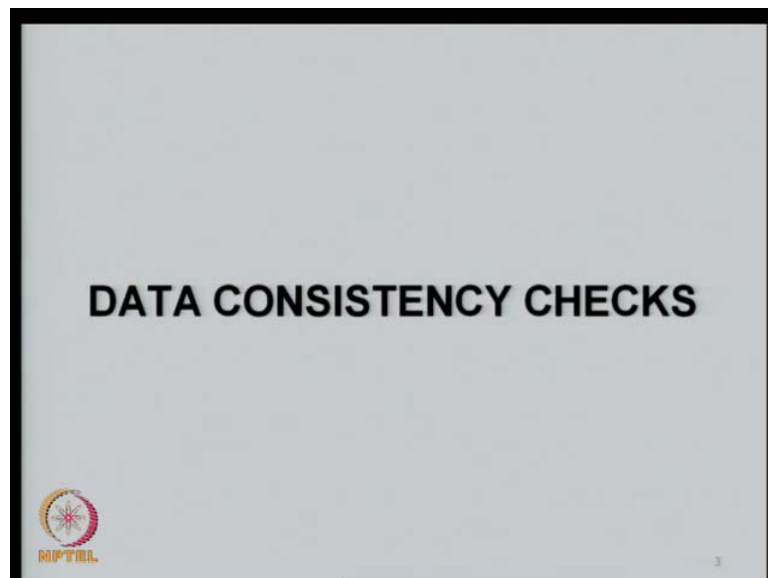
Now, we also determined the coefficients a and b . We gave a as $m_1 m_0^{-1}$ and b as $m_0^{-1/2} (m_0 - m_1 m_0^{-1} m_1^T)^{1/2}$ where m_0 is a cross correlation matrix of lag zero and m_1 is a cross correlation of lag one and the p here is the number of sites.

So, we are typically talking about simultaneous generation of flow data at various locations. With this, now we are concluding the theoretical aspects of stochastic hydrology. So, essentially we have developed several models. We have discussed several models for flow generation, we also introduced the conditional probabilities and so on.

So, the complete theoretical aspect of stochastic hydrology has been covered so far. What we will do now onwards is to look at some practical aspects and some applications. Towards the end of the course I may also talk about the climate change aspect and the non stationery introduced by climate change and so on.

So, now onwards the focus of the course will not be so much on the theoretical aspects, but, on the practical aspects and how we apply all the knowledge and the **and the** coverage, the theoretical coverage that we have developed so far in this course to some practical problems.

(Refer Slide Time: 03:38)



As you can see or as you can recall most of the problems that we have dealt with in this course stochastic hydrology are all data dependent. In the sense that you have historical or observed data or assimilated data and all the analysis that you have done in this course is mainly on the data and therefore, when you want to apply this techniques to practical problems.

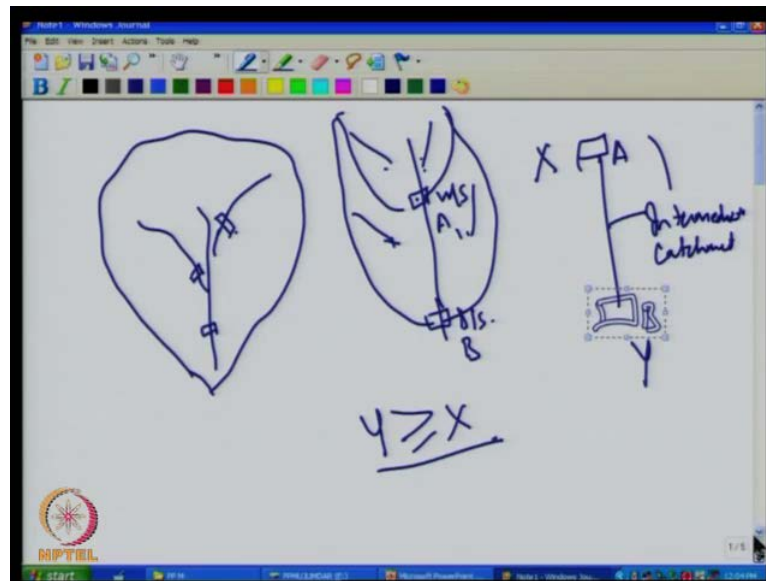
Let us say that you have a catchment which has to develop for water resources development through putting certain reservoirs or hydro power plants or lift irrigation projects and so on. The hydrology of that particular basin must be analyzed with the techniques that we have studied in this particular course.

So, the data analysis becomes extremely important. It is not the analysis in the sense that we apply the techniques that we have studied in this particular course not that. Before even going to the techniques that we have discussed, we need to do certain checks on the data how much data is available at what location, what kind of data that we need to use out of the data that is available and required for these models and with respect to the data that is observed elsewhere in the same basin whether the data that we are taking at a particular location is consistent and whether there have been any major changes that have occurred in the basin that have affected the data which are shown as signals in the data itself.

So, in this lecture and perhaps part of the next lecture what we will do is we will look at how to get the information from the data in terms of the consistency of the data that we are seeing with respect to the other gauges in the basin.

So, for the time being now, we will deviate slightly from the rigorous theoretical mathematical aspects of stochastic hydrology to the more mundane yet extremely important aspects of data consistency. So, the problem that we will be discussing now is the following.

(Refer Slide Time: 06:22)



Let us say that you have a basin like this and then you have a major stream coming here and then there may be some tributaries and so on. You may have data at this location **you may have data at this location** what I mean by that is that you may have a stream gauge at this location, stream gauge at this location, stream gauge at this location and so on.

So, you have the observed flows let say for the last thirty years at this location, this location and so on. Just look at two gauges one an upstream gauge and another a downstream gauge. The flows that are recorded at this location the upstream gauge will be the flows that are generated from the catchment upstream of that and therefore, the entire catchment that is contributing to this gauge is recorded by the entire catchment that is contributing to this location is recorded by the gauge that is located here. When we come down and look at this gauge here, this will record all the flows that have come from the entire catchment upstream of this particular gauge.

Let say we call it as gauge A and this is gauge B. So, the flows that are recorded at gauge b will also include the flows that are recorded at gauge A assuming that there are no control structure. By control structures I mean let say you have a bear or a reservoir and so on which will regulate the flow. So, they are assuming that between a and b and even upstream of a there are no control structures. So, we are only talking about the natural flows.

When you have the data at a as well as at b for concurrent periods; let say that we have thirty years of data, thirty years of monthly data is what we are examining. So, when you have data at A and B during concurrent periods; obviously, the flows at B must be greater than or equal to the flows at A. Why they should be greater? Because from the flows at A you also add the intermediate catchment flows.

So, the catchment that is upstream of b and downstream of a which is not included in the catchment of a; this catchment also contributes to the flows at b and therefore, the flow at b will be greater than or equal to flow at a provided there is no significant or there is no utilization of the flows **in in** the intervening catchment.

The catchment between this gauge a and the gauge b is called as a intermediate catchment. Now, specifically in our country you know we may have data for about 30 to 40 years. If the system is well developed perhaps we may have even more number of years. But typically we may have data for about thirty to forty years and in many cases we will have data for much smaller periods 15 years 12 years and so on. But, we have to use this available data. We cannot simply keep quite saying that the data is. So, small and therefore, we would not be able to do anything.

Whatever data that is available, we must be able to use in a meaningful manner and then make decisions about water resources development and therefore, all the hydrologic analysis that we have studied so far will become useful only if we understand the data issues correctly and then do the checks that I am going to discuss now before we go into the stochastic hydrology aspects **alright**.

So, if you have two gauges like this A and B; let say I denote the flow at b as y and the flow at a as x. Now, these may be vectors or matrices depending on the way you are in the time series. So, this is y and this is x. Y has to be greater than or equal to X. This is a first observation. So, which means that you are given data at these two gauges the first observation that you would like to make is the flow that you have observed at y at v which is namely y must be greater than equal to the flow at gauge a which is x.

If you say, if you see or observe that this in fact is not valid that is y is not greater than equal to x, but, y is less than equal to x less than x perhaps. What does that imply? It implies that the flow that was observed here has been reduced by the time it came here. If you have gauges and the situation that y is less than equal to x, less than x. In fact, you

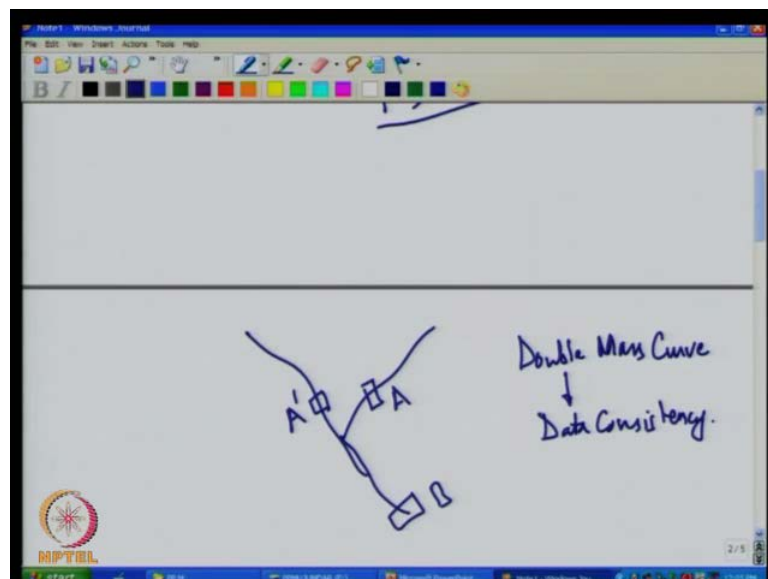
should then look at what is happening in the intermediate catchment. Is there a significant utilization of water that we are not accounted for?

Let us say there is a major lift irrigation project which is taking out water that is coming from here and the intermediate catchment is receiving less rain and therefore, the intermediate catchment contribution itself is not so high as to compensate for the water that is utilized through the lift irrigation projects and so on.

So, the moment you see that y is less than or equal to x ; immediately you have to look at what are the activities that are taking place between A and B in terms of water utilization. So, y will be less than or equal to x only if the rainfall is significantly smaller compared to the rainfall upstream of this rain gauge A or and perhaps and the utilization in the intermediate catchment is significantly high.

So, you must immediately be alerted to the situation of utilization as well as the amount of rainfall in the intermediate catchment. Now, when I say y is greater than equal to x it is not that individual values you keep on checking. So, there must be ways of checking that y is in fact greater than equal to x or not. So, this is what we do through data consistency checks.

(Refer Slide Time: 13:40)



So, the first level of data consistency check that we do is to examine the flows at this location, at a particular location with respect to the flows at all the upstream location.

What I mean by all the upstream location is that, you may be focusing on a particular gauge. Let us say that the gauge is here and this is a stream and there may be one gauge here, there may be another gauge here and then this is a gauge that we are talking about. Let us say this is a gauge b and this is a and a dash.

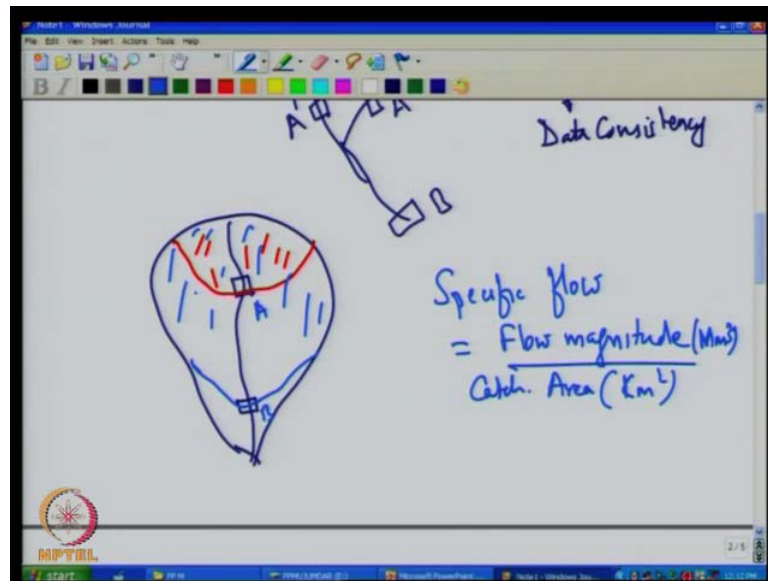
So, the flow at b must be consistent in terms of its magnitude with respect to the flow at A as well as with respect to the flow at A dash. This is what we do through what is called as a double mass curve. This is a first level of data consistency check that we do. (Refer Slide Time: 06:22)

Now, remember when we are doing a data consistency for any basin like this; we are assuming a hydrologic homogeneity in the sense that the runoff generating mechanism at each of these locations is the same, which means the land used pattern is the same, the rainfall pattern is the same etc and then we do this data consistency checks.

If everything is same I mean if the region is hydrologically homogeneous and yet we do not get such a situation then it means two or three things; one is as I said the intermediate catchment rainfall itself may be very small, there may a significant water utilization which we are not accounted for **in in** in the data that is obtained at B or there may be something wrong with the data that is observed at either b or a. The last point is what we must be alerted to that is what indicates data inconsistency. Now, this data inconsistency may arise because of some changes that are taken place in the catchment.

Let us say suddenly you start measuring with a different gauge altogether or there may have been some sudden changes that have occurred in the catchment which have affected the flows at b and so on. So, it indicates inconsistency of data at b with respect to the data that is observed upstream in the upstream gauges.

(Refer Slide Time: 16:16)



We will also see one other aspect now for the data consistency. Let us say that you have a fully hydrologically homogenous region. By hydrologically homogeneous region I mean the runoff generation capacity, runoff generation mechanism which means what? The slope show the catchment characteristics in terms of the slope, in terms of the vegetation the land used patterns etc are uniform all through the watershed that we are considering all through the basin that we are considering.

If that is so; then let say that we have gauges like this we have a rain gauge like this and we have a rain gauge like this. Now, this rain gauge will account for or will measure assume that the flows that are governed, flows that are generated in this particular basin is measured at this gauge and this gauge here will measure the flow that is generated at all of this including this part. So, this is rain gauge A and this is rain gauge B.

Let us say that we are looking at the annual flows at this point. Let us say some 5 hundred million cubic meters of flow that has taken place at this at this location in the entire year. Now, this flow has taken place because of the catchment that is marked in red here. So, all of this catchment has contributed to the flow at that location for a particular annual rainfall. So, that annual rainfall has contributed to flow at this location and that is some 500 million cubic meters or something.

If you divide that flow annual flow by the area; we get what is called as a specific flow. So, we will say the flow magnitude divided by the area that is catchment area. Let us say

this is million cubic meters and this is in square kilometers. So, typically the specific flow that we calculate will be in million cubic meters per square kilometers. So, So, if you convert the units etc we may get the depth of flow **depth of flow** in let say meters or some such depth units.

If your catchment is in fact, homogeneous then the specific flow that you calculate at this location a and the specific flow that you calculate at b must be same if it is perfectly homogeneous in terms of rainfall. That means, the rainfall has occurred uniformly all through the year across the complete catchment and the runoff generation mechanism is completely uniform in terms of its land used in terms of the soil in terms of the antecedent moisture contained and in terms of the vegetation etc. All the catchment characteristics are the same across all through the catchment.

In which case because you are putting the same rainfall at b as well as at a and because the runoff generation mechanism is the same for this catchment shown in red and this entire catchment shown in blue; the specific flow here must be the same as the specific flow here. This is another consistency check that we do.

However, the catchment characteristics in practical situations will not be exactly same and also the rainfall of distribution may not be exactly the same and therefore, what we do is the specific flow at b must be comparable with specific flow at a is the criteria that **we criterion that** we use for checking the consistency.

So, the two levels of consistency that we can do is one look at a particular gauge and then look at its consistency with respect to the gauges upstream of it in terms of just the magnitudes of flow and this we do by what is called as the double mass curve which I will introduce just now. The other one is that you look at the specific flows which is essentially flow per unit area and that flow per unit area must be comparable across different gauges in the same site.

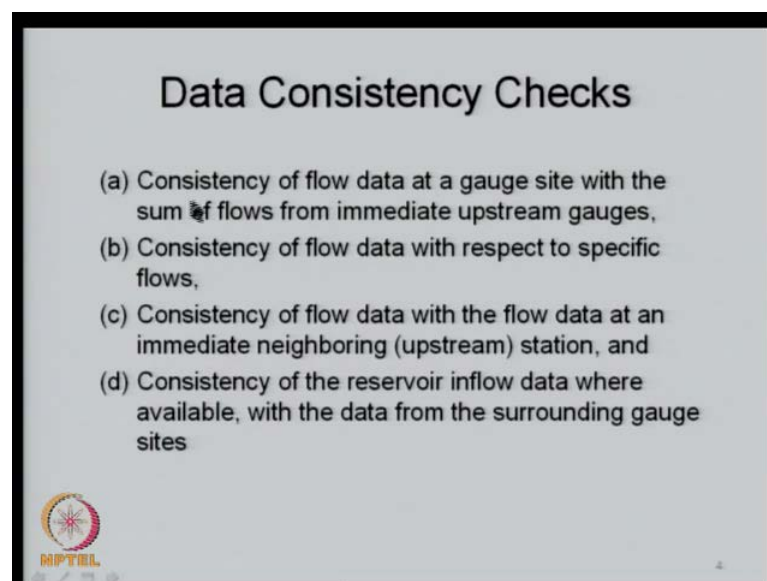
Now, let us say in this case the specific flow at b at this location at b is. In fact, not the same as a, specific flow at a. Again as I explained in terms of for your magnitudes what that may indicate is let us say the specific flow at this location is much smaller compared to specific flow at a then it may indicate a significant utilization of water here in this intermediate catchment between b and a or a significantly lower rainfall in this

catchment or the catchment characteristics between b and a are much different from the catchment characteristics at this **this** catchment a.

A combination of all of these factors or all of them together and therefore, whenever you look at the specific flow at a particular gauge and it shows some inconsistency. You must look at what is happening in the **what is happening to the** physical characteristics of the catchment as well as to the hydrologic characteristics in terms of the rainfall as well as in terms of the runoff generating mechanisms.

So, this is what we do before we proceed to using any analytical techniques that we have covered in the stochastic hydrology course and these are specs of the data are often much more important than the theoretical aspects because the theoretical aspects can be done rigorously because the methods are all very clear. But, they are all based on the data and therefore, the data consistency checks must be done for any large scale hydrologic analysis.

(Refer Slide Time: 23:40)



So, we will do the, we will now introduce the data consistency checks more formally and examine with respect to some examples. So, as I mentioned first we look at consistency of flow data at a gauge site with the sum of flows from immediate upstream gauges. The y which is observed at that particular location must be greater than equal to x which is the total flow that is coming up to the upstream gauges. Then, consistency of flow data with respect to specific flows; So, you look at the specific flows at a particular location,

in the same basin. This must be consistent in terms of it being comparable to the specific flow at other locations in the same basin provided the basin is more or less homogeneous. Then consistency of flow data with the flow data at an immediate neighboring upstream station; So, one is look at all the upstream gauge station and then look at the consistency. Another is you may have a neighboring station in the same hydrologically homogeneous region then you look at the consistency between these stations. So, the two station data must be consistent.

Then often we have reservoir sites. Now, the reservoir sites have their own inflows recorded. Now, these inflows must be consistent with neighboring gauges or surrounding gauges in the same hydrologic region. So, these are typically the consistency checks that we do on the data before we proceed to any hydrologic data analysis.

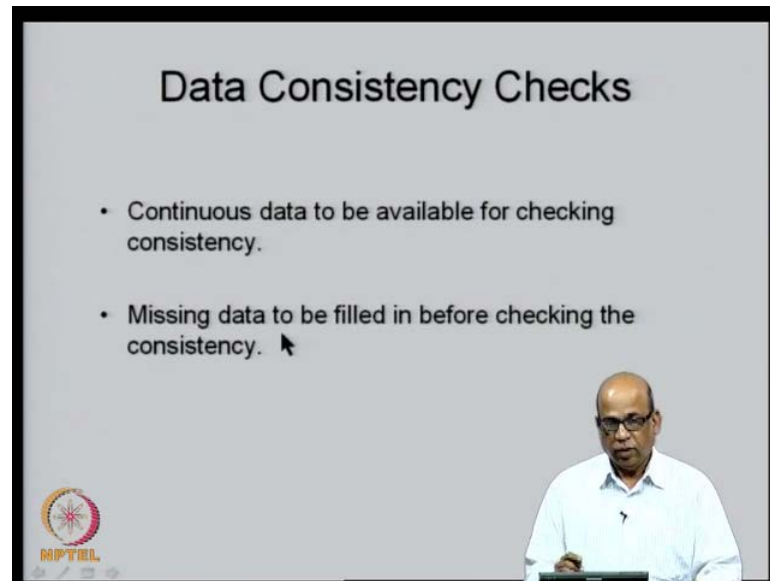
Typically in our country most specifically in our country, we may have data for a short duration. Let us say that there are major water resources structures existing for which you want to do analysis in terms of water utilization and so on. However, the data that is available may be of the order of 5 years ten years 12 years and so on which for any meaningful exercise may appear to be absolutely inadequate. However, we need to use this data and do the best that is possible from the available information and the data.

A second level complication is even in this limited data or even if the data is available for larger length; there will be missing data. In the sense that either the data has not been properly recorded or the data has not been recorded at all and therefore, we must first look at the continuity of the data and what do we do for those periods in which there is absolutely no data available. So, even before we go to the consistency checks we must feel the missing data. So, the first step that we do in any hydrologic analysis is to look at the time series of the observed data and look at the breaks that happened in the time series for several reasons.

One is the data itself may not be recorded or the data may be wrongly recorded and therefore, it has been taken out of the record or there may be some gaps or breaks that cannot be explained in the time series and therefore, the breaks have happened. So, the time series of data has breaks in between. Now, these breaks can be just a point break in terms of, if you have a daily time series only one day data is missing or it may be continuous. Let us say 3 or 4 days continuously the data is missing and so on.

So, we must first look at some methods by which we fill this missing data and then complete the time series. Now, the filling of the data is not very rigorous. It does not include very rigorous mathematical procedures. But, it includes certain rather empirical procedures which use judgment and which will make use of the fact that we have the data must be consistent with the other data in the time series.

(Refer Slide Time: 28:04)



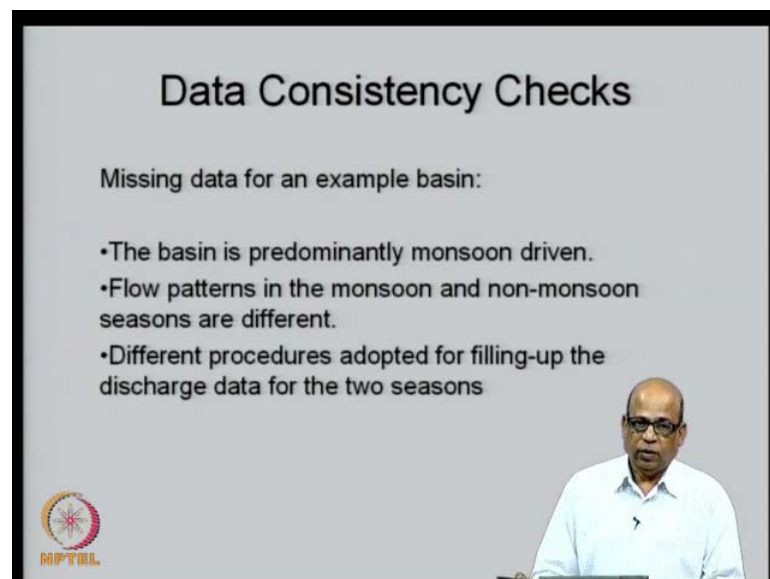
Data Consistency Checks

- Continuous data to be available for checking consistency.
- Missing data to be filled in before checking the consistency. ↖

MPTEL

The slide features a presenter in a light blue shirt and glasses in the bottom right corner. The MPTEL logo is in the bottom left corner.

(Refer Slide Time: 28:14)



Data Consistency Checks

Missing data for an example basin:

- The basin is predominantly monsoon driven.
- Flow patterns in the monsoon and non-monsoon seasons are different.
- Different procedures adopted for filling-up the discharge data for the two seasons

MPTEL

The slide features a presenter in a light blue shirt and glasses in the bottom right corner. The MPTEL logo is in the bottom left corner.

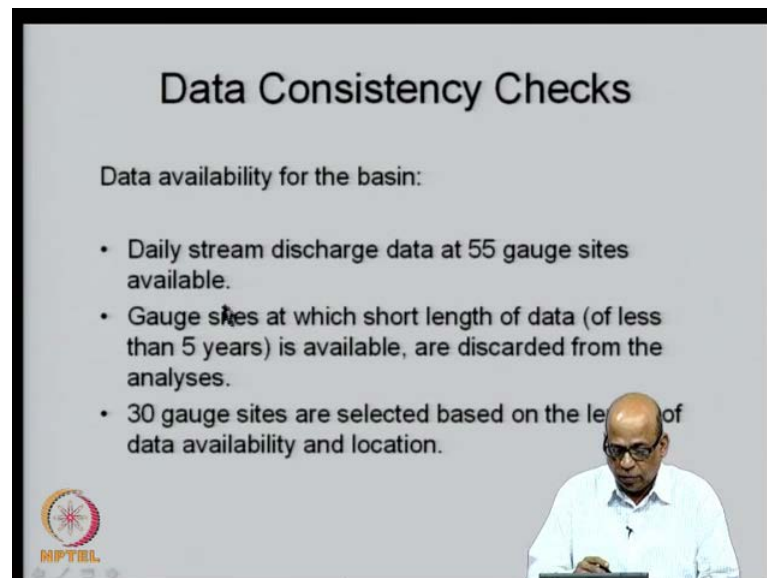
So, let us look at some simple methods by which we fill the data now. So, the missing data should be filled in even before we attempt the data consistency. We take an example

basin as I said you know the missing data the filling up procedures will be mostly judgmental, mostly based on judgment and therefore, you look at the time series for that particular basin and then see what type of data that is available and within that data how much is missing, at what time locations it is missing and what are the ways by which we can fill in the data.

So, the first thing that we do is; we look for this example basin. We see that the flow patterns in monsoon and non monsoon periods are much different. So, the monsoon flows are consistently higher, there are very high discharges that are taking place whereas, the non monsoon flows are essentially because of the base flow and so on.

So, the patterns are different and therefore, first level is that you look at the missing data in the monsoon region, you adopt a different procedure for **monsoon** monsoon periods and a different procedure for non monsoon periods.

(Refer Slide Time: 29:34)



Data Consistency Checks

Data availability for the basin:

- Daily stream discharge data at 55 gauge sites available.
- Gauge sites at which short length of data (of less than 5 years) is available, are discarded from the analyses.
- 30 gauge sites are selected based on the level of data availability and location.

MPTEL

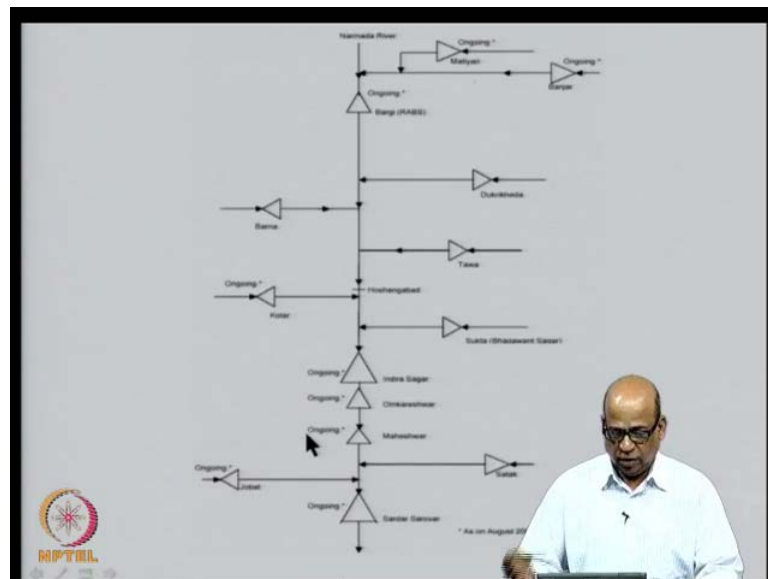
The slide features a presenter in the bottom right corner, wearing a light blue shirt and glasses, looking down at a laptop. The MPTEL logo is visible in the bottom left corner of the slide.

So, for the monsoon period for this example basin let us say I will show you the basin also. **It is a**, We have daily discharged data for 55 rain gauge sites the stream low stream gauge sites. So, at 55 sites we have the data available, but, many of which will have very short length of data. Let us say of less than 5 years. There have been just the gauges have been just put in place or even if they are put in place earlier data has not been recorded or data has not been filed and therefore, we have very short length of data of less than 5 years and so on.

Now, because we have a large number of gauge sites; we can afford to ignore these gauges at which the data is very **very** small. So, we just discard data of length less than 5 years. Then out of 55 sites, we take out about 25 sites and then take only about 30 sites.

So, whenever you have large catchments, large water resource systems where a number of gauges exist, first look at the length of the data that is available at each of the gauge sites and relatively unimportant gauges which have very short length of data must be discarded from the analysis. Otherwise they will introduce a very significant bias in the analysis and because of the short duration of the data and therefore, these must be taken out.

(Refer Slide Time: 31:16)

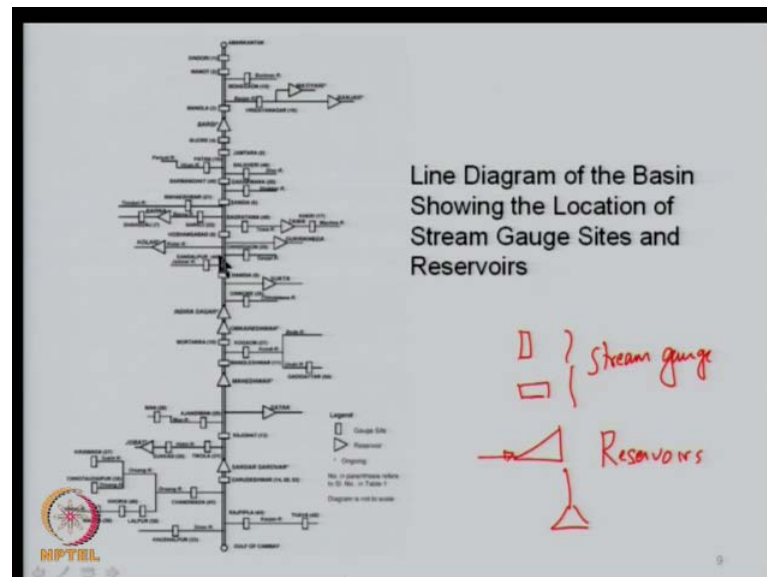


So, this is a system it is. In fact, the Narmada system where there are large number of reservoirs that are planned or been put in place or already in place. This is Bargi for example, this is Indirasagar, this is Sardar Sarovar and there are other major reservoirs like Matiyari is here, Kolar barna and so on.

So, this is a major reservoir system and the Madhya Pradesh border end somewhere here just before the Sardar Sarovar as all of us Indians know that this is an extremely important water resource system and therefore, the hydrologic analysis for this system is important. However, there may not be enough data, adequate data available at all the locations, but, luckily for Narmada system there is significant data that is available at several locations because of the importance of this system.

So, at least about 20 to 30 years of data is available at major locations and in fact, reconstructed data is available for much longer length and so on. So, this is a good system to analyze and also the Narmada control authority also maintains **record** records of the data in a very good manner and therefore, it is available for analysis.

(Refer Slide Time: 32:34)



(Refer Slide Time: 31:16)

You look at the gauge sites now. This is your system and at several locations there are stream gauge sites. So, this is the stream gauge location. The marks like this, like this the rectangular marks either this way or this way; these indicate the stream gauges and the reservoirs are located or denoted like this these are this way or this way these are the reservoirs. So, these are the stream gauges at these gauges the data will be available stream flow data will be available.

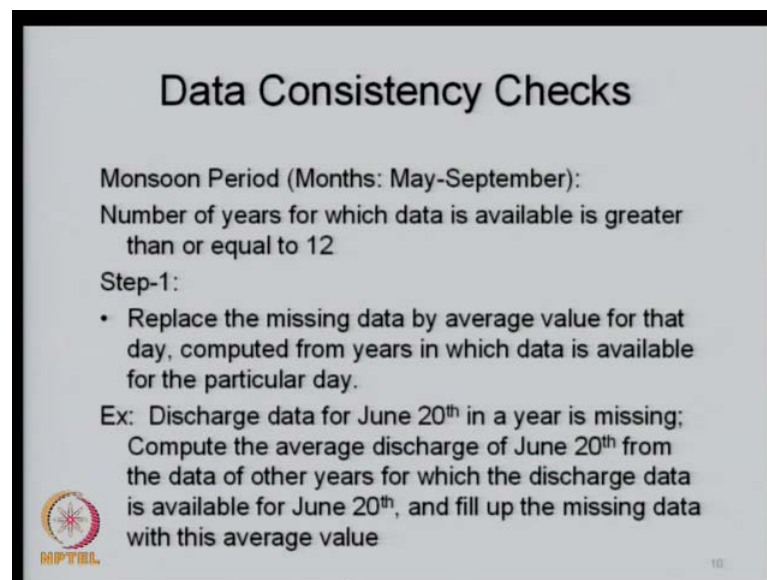
So, there are a large number of stream gauges, but, there are certain important stream gauges. For example, you take Rajghat which will measure what is coming into Sardar Sarovar on the main river whether or also Tikola and Sukad which are adding to the Sardar Sarovar reservoir here.

Similarly, you look at Maheshwar, Omkareswar, Narmada Sagar etc; the gauge at Handia will be measuring the flows that are coming into this. So, if we look at gauges here in the main stream, let us say focus on one of the gauges Handia here. Let us say the

Handia gauges here. Immediate upstream of that is a gauge at Hoshangabad and for Hoshangabad there is a Sandia gauge.

So, if you look at only the mean stream there are several gauges which are located between the reservoir Bargi and the Narmada sagar which is a large intermediate catchment. So, we will just look at how do we do the data consistency checks for such a configuration of stream gauges, a large number of stream gauges at all of which we have the historical data available.

(Refer Slide Time: 35:16)




Data Consistency Checks

Monsoon Period (Months: May-September):
Number of years for which data is available is greater than or equal to 12

Step-1:

- Replace the missing data by average value for that day, computed from years in which data is available for the particular day.

Ex: Discharge data for June 20th in a year is missing; Compute the average discharge of June 20th from the data of other years for which the discharge data is available for June 20th, and fill up the missing data with this average value

 NPTEL 10

So, we will just examine the type of data consistency checks that we do here. So, as I said because the flow generation mechanisms are different in monsoon and non monsoon periods; we do this separately for the monsoon period as well as the non monsoon period. So, first we choose those particular gauges where at least 12 years of data is available. Now, this 12 is an arbitrary number because we have discarded the rain gauges the stream gauges which are less than 5 years of data. 12 years is the reasonable length of data.

So, we choose only those gauges which have 12 years of data first as a first step. So, those gauges which have at least 12 years of data what we do is we replace the missing data by average value for that day. This is a daily data we are talking about.

So, for that day you take the average of the data from all the remaining years. Let us say a particular year June 20th is missing. Then June 20th average year average flow for all the remaining years you take and then simply replaces by that very simple procedure.

However, when we do this you re plot the time series and then see that it is a smooth time series with respect to the 19th of June flow, 21st of June flow etc. It should not be much more different and it should not introduce a non smoothness in the time series.

(Refer Slide Time: 36:56)

Data Consistency Checks

Step-2:

- Plot the time series of daily flows and check whether there are any significant fluctuations in the hydrograph. If such a case is noticed, correct the filled up data to smoothen the hydrograph.

The slide contains two hand-drawn hydrographs. The left one shows a blue line with a sharp peak and a red line with a smoother curve. The right one shows a blue line with a sharp peak and a red line with a smoother curve, with a red circle around the peak area.

IPTCL

11

So, the first level that we do is simply take the average and put it there. When we do that like I said we plot the time series again and look at the surrounding thing. Let us say this was your time series and this is a missing data that we are using. Let us say we put the missing data here. So, it goes up here and then it comes down here. Let say some 3, 4 days continuously we are filling it up.

Now, this should be a smooth curve like this. So, this is the remaining period and these are the missing data that we have filled in using the average is for those days. So, what we have done? For this day we put it as a average flow computed from the remaining years for the same day and then similarly, if this was missing we do the same thing for that day and so on.

Then make sure that the time series continues in a smooth manner. Let us say the time series did not continue in a smooth manner. Let us say that this was your time series and

then you started filling the data and then you saw that the filled data comes somewhere here, like this with the average flows and then the time series continues at this level again like this.

So, essentially what we did in filling the data is that we have disturbed the smoothness of the time series from here to here. And then it starts coming at this location. When that happens; it means that the procedure did not work out well. So, like this it happened and then it came back to this point and all of this is the missing data.

(Refer Slide Time: 38:58)

Data Consistency Checks

Monsoon Period (Months: May-September):

Number of years for which data is in between 5 and 12

- If the missing data is non continuous (i.e., not more than 3 days continuously in a month), plot the daily flow time series for the month and join the curve smoothly to obtain the missing values

The slide includes a logo for NPTEL in the bottom left corner and a small number '12' in the bottom right corner. Handwritten red ink diagrams illustrate the concept: one shows a jagged line with a gap labeled 'June' and 'July', and another shows a smooth curve with a gap labeled 'June' and 'July'.

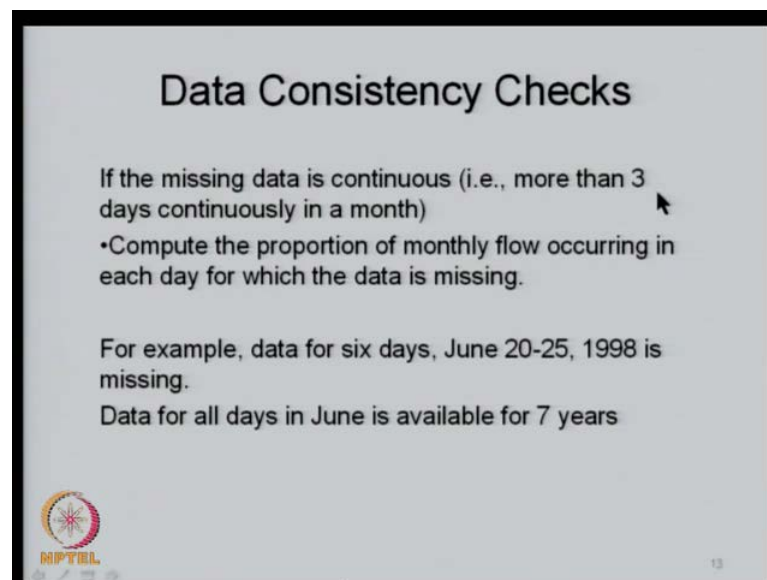
So, if this happens then the procedure of missing data that we have used is not good and then we have to go for some other methods which I will discuss now. Then in the monsoon periods, the second case is let say that we have data between 5 years and 12 years. What we did for this example is that anything less than 5 we have discarded and anything more than 12 we just introduced one of the methods and then examined whether the time series will be smooth by the method that we have used.

Now, we will see what we do between 5 and for those gauges which have data between 5 years and 12 years. Let us say if the missing data is non continuous. What do I mean by non continuous? That is if we have data like this, this was a time series and then one day is missing here and then it is continuous. So, there is only one day that is missing here or may be maximum of 3 days is missing here like this. Three days data is missing and then

continues. So, if the missing data is non-continuous then plot the daily flow time series for the month and join the curves smoothly to obtain the missing values. That is all.

So, let say this was June month for so **so** many years for June month we have the flow data for. So, many years simply plot the time series and then smoothly join it if you just put three days you can smoothly join it because the data itself is small the length of the data itself is small and therefore, the procedure that you adopt can be slightly approximate.

(Refer Slide Time: 40:44)




Data Consistency Checks

If the missing data is continuous (i.e., more than 3 days continuously in a month)

- Compute the proportion of monthly flow occurring in each day for which the data is missing.

For example, data for six days, June 20-25, 1998 is missing.

Data for all days in June is available for 7 years

 NPTEL

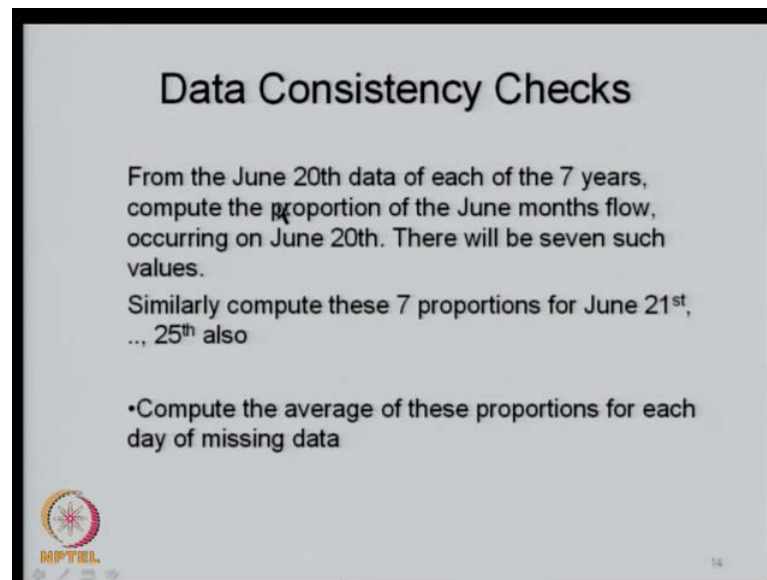
13

So, you simply smoothly join the data, join the time series and then use them as the missing data. Now, if the missing data is continuous; that means, more than 3 to 4 days continuously in a month. Let us say June month; 10 days is missing then we adopt slightly different method here. What we do? We compute the proportion of monthly flow occurring in each day for which the data is missing.

(Refer Slide Time: 40:44)

Let me explain this. These are all very simple procedures, but, you know they are extremely useful. Let us say June 20th to 25th the data is missing for a particular year. Then data for all days in June is available for 7 years because you are talking about the case where 5 to 12 years data only is available. So, 7 years data is available and in that a particular year June 20th to 25th is missing.

(Refer Slide Time: 41:40)




Data Consistency Checks

From the June 20th data of each of the 7 years, compute the proportion of the June months flow, occurring on June 20th. There will be seven such values.

Similarly compute these 7 proportions for June 21st, ... 25th also

- Compute the average of these proportions for each day of missing data

 14

Then what we do? From the June 20th data of each of the 7 years that is available data we compute the proportions of the June month's flow occurring on the June 20th. This is a date on which the data is missing. There will be 7 such values.

(Refer Slide Time: 40:44)

(Refer Slide Time: 41:40)

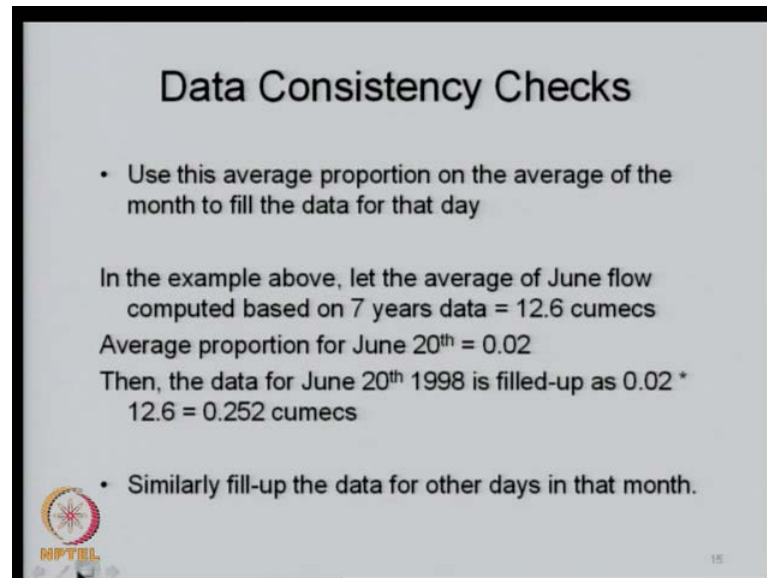
So, similarly, compute this 7 proportions for June 21st to 25th also. What we mean by this is you just so that you do not miss the point I repeat this. We have data for all days in June for 7 years which means we have 7 years of data where nothing is missing and we use that and then get the proportion of these flows at this dates.

So, this we do for all of the other days on which it is missing also. Then compute the average of these proportions for each day of missing data. So, we computed the proportion. By proportion what I mean by that is that you have a June flow, total June flow and in that we have 21st 22nd etc and these values are available for 7 years.

So, let say this was x and this total flow $x \cdot 1$ and this total flow was y . So, we get $x \cdot 1$ by y $x \cdot 2$ by y and so on. Like this we get for all the 7 years, each of the 7 years and then take the average of these proportions. So, compute the average of these proportions for each days of missing. So, you have 7 values you take the average of this then use this average proportion on the average of the month to fill the day that is all.

So, we had for this 1998, we had 7 days this 21st to 25th is missing. So, for 21st we compute the average and knowing the total flow for 1998 June, we use this proportion and fill the data. So, that is how we fill the data for that.

(Refer Slide Time: 44:04)




Data Consistency Checks

- Use this average proportion on the average of the month to fill the data for that day

In the example above, let the average of June flow computed based on 7 years data = 12.6 cumecs
Average proportion for June 20th = 0.02
Then, the data for June 20th 1998 is filled-up as $0.02 * 12.6 = 0.252$ cumecs

- Similarly fill-up the data for other days in that month.

 15

So, in the example above, just to give this a complete picture we take the average of June flow computed based on 7 years of data is 12.6 six cubic meters per second. Then average proportion for June 20th is 0.02. What is this 0.02?

This is what we computed based on all the available data for June 20th with respect to the total flow. Data for June 30th 1998 is 0.02 into 12.6. This is for 12.6 June month **june month** 0.252 cubic meters per second. So, like this we fill up for individual days for each of the data.

(Refer Slide Time: 44:54)


Data Consistency Checks

Non-monsoon Period (Months: January-April and October-December):

- Compute the probability that the flow is non-zero for that day. (Clarke, 1973)

Ex: Day - Oct.20th; Number of years of data available = 7

Number of non-zero flows on Oct. 20th in the 6 years (leaving out the year in which the data is being filled) = 2 (i.e., only 2 out of these 6 years have a non-zero flow. Remaining 4 years have a zero flow)



Clarke, R.T. (1973), Mathematical Models in Hydrology, FAO Irrigation and Drainage Paper no. 19, Rome, Italy.

18

Now, we go to non monsoon period. Now, in the non monsoon period you may have a rainfall or you may you may have a flow or you may not have any flow at all which means if you are talking about small distributaries and so on small tributaries you may not have any flows at all in those rivers.

So, we compute the probability that the flow is non zero for that day. Just look at the historical data and then get these probabilities. Once you know that it is a non zero; that means, you have the probability of non zero flow for a particular day.

Let say example is October 20th number of years of data available is 7 and number of flows non zero flows **in** in the 6 year period leaving out the year in which the data is being filled. So, you compute the proportion. I will not call it as probability, but, proportion of the number of periods in which the flow is non zero for that particular period.

(Refer Slide Time: 46:12)

Data Consistency Checks

Probability that a non-zero flow occurs on Oct. 20th = $2/6 = 0.333$

- Generate a uniform random number between 0 and 1

e.g., for the Oct. 20th missing data in (a) ab let this random number be 0.248

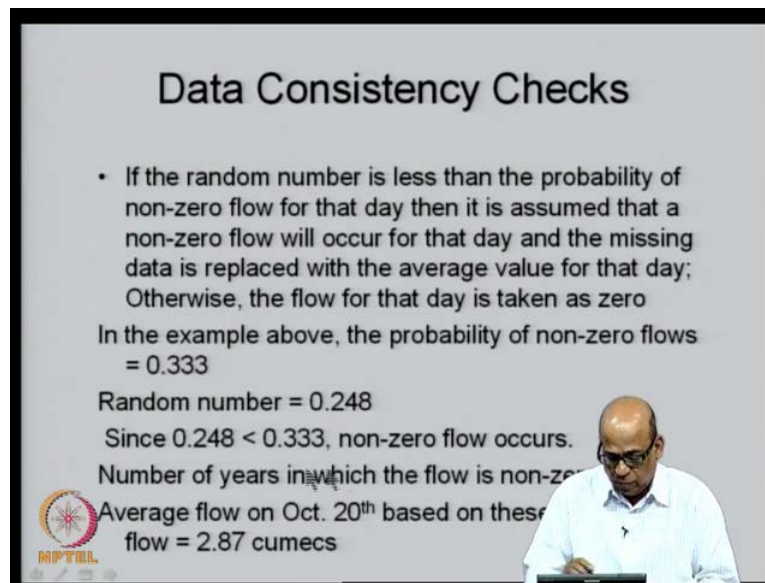
MPTEL

Let us say we get this probability as 0.333 in this particular case. So, for a particular day we know the estimate of the probability of flow being non zero. Now, we generate a random number. Simply you take a uniform random number from your calculator or otherwise you just get the uniformly distributed random number.

If this random number is smaller than this probability, you say that it is a non zero flow or if it is more than this then you say it is a zero flow. Now, this can be interchanged because you are generating a uniformly distributed random number which can be either smaller than this or higher than this. Just be consistent that is all. That means, if it is less than if the random number is less than this probability I will call it as a non zero flow.

If it is more than this number then I will call it as a zero flow. It can be other way round also because you have to generate some number of random numbers. So, for the 20th missing data October 20th missing data let this random number be 0.248 if it is 0.248 and they are following the convention that if it is smaller than that I will call it as a non zero flow.

(Refer Slide Time: 47:34)



Data Consistency Checks

- If the random number is less than the probability of non-zero flow for that day then it is assumed that a non-zero flow will occur for that day and the missing data is replaced with the average value for that day; Otherwise, the flow for that day is taken as zero



In the example above, the probability of non-zero flows = 0.333

Random number = 0.248

Since $0.248 < 0.333$, non-zero flow occurs.

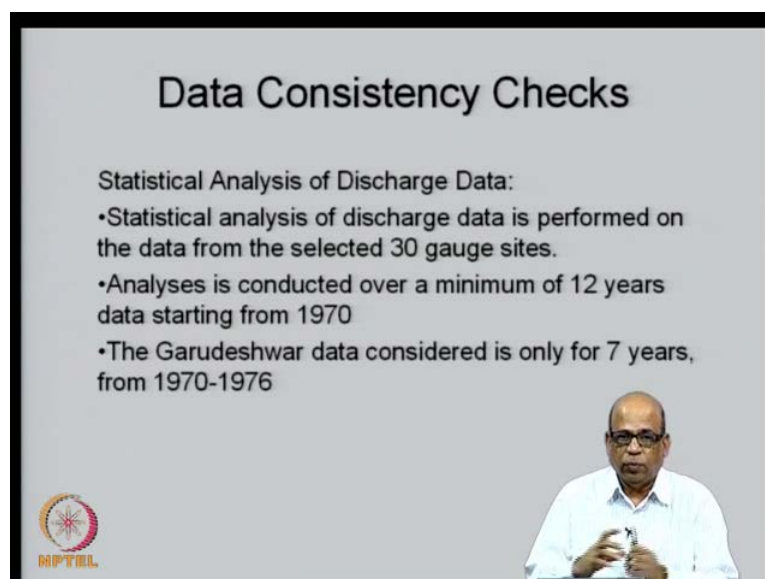
Number of years in which the flow is non-zero = 10

Average flow on Oct. 20th based on these 10 years = 2.87 cumecs



So, we say this is a non zero flow. **So, we say this is a non zero flow.** So, if the random number is less than the probability of non-zero flow this is a procedure that is given by and it has been quite useful. So, this will be a non-zero flow. Once you identify that it is a non zero flow which means that the flow is in fact occurring then you just use the average flow as we have done earlier for that particular day and then fill it with the average flow.



(Refer Slide Time: 48:05)



Data Consistency Checks

Statistical Analysis of Discharge Data:

- Statistical analysis of discharge data is performed on the data from the selected 30 gauge sites.
- Analyses is conducted over a minimum of 12 years data starting from 1970
- The Garudeshwar data considered is only for 7 years, from 1970-1976



So, first decide whether it is a zero flow or non-zero flow. If it is non-zero flow simply fill it by the average of that flow. Once you fill the data missing data then you have the complete time series ready with you.

So, this time series now we will subject to consistency checks now. So, the first step that we did for the data analysis before going to any stochastic models is to look at the time series that is observed at each of the locations and make the time series continuous and complete by filling the missing data. So, missing data filling is an important although a simple in terms of the mathematical rigor it is a very simple and perhaps empirical way of doing it. However, this is an important exercise. So, that you have a complete and continuous time series available at each of the locations.

Then we ask the question; we have several such gauging locations and we have at each of these gauging locations, we have a reasonable length of data of about 12 years or something. Then we will start looking at what is the consistency of the data at a particular location (()) the data at other locations how consistent is **consistent is** the data at a particular location with respect to other locations in some statistical sense.

(Refer Slide Time: 49:40)

Statistics of Annual Flows

S.No	Gauge site	Data used (Period)	Duration (years)	Average (MCum)	Maximum daily flow (cumec)	Maximum (MCum)	Minimum (MCum)	Standard deviation (MCum)	Coeff. of variation (%)
1	Dindori	1988-1999	12	1,251.50	2,624.00	2,519.00	760.19	438.13	35
2	Manot	1976-1999	24	3,042.13	6,180.00	5,427.17	1,012.56	988.13	32
3	Mandia Town	1977-1980 1993-1995	7	4,588.07	8,409.71	14,336.93	890.92	4,594.22	100
4	Bijore	1988-1999	12	8,713.39	20,349.00	16,473.11	2,870.88	4,168.42	48
5	Jamtara	1971-1999	29	9,179.34	21,355.10	20,985.05	2,371.81	3,706.08	40
6	Sandia	1978-1999	22	14,419.68	18,160.00	37,623.47	6,066.00	6,744.77	47
7	Hoshangabad	1972-1999	28	22,932.10	31,600.00	53,146.10	8,052.22	10,279.85	45
8	Handia	1977-1999	23	25,304.99	26,240.00	60,253.63	11,415.29	10,615.42	42
9	Mortakka	1970-1978 1988-1999	21	31,300.49	59,371.49	62,923.09	16,158.82	13,087.86	42
10	Mandleshwar	1971-1999	29	33,239.03	100,096.80	69,615.62	15,226.93	13,600.32	41
11	Rajghat	1971-1999	29	33,777.94	56,601.30	74,077.80	14,951.77	13,858.23	41
12	Garudeshwar	1971-1975 1980-1999	25	32,254.93	53,749.00	74,077.70	15,513.25	14,016.17	43
13	Mohegaon	1977-1999	23	2,199.34	6,526.90	4,240.24	646.21	791.37	36
14	Shridayanagar	1976-1999	24	1,549.61	3,632.00	4,756.16	544.39	891.54	58
15	Patan	1979-1999	21	1,538.53	1,817.00	3,552.23	497.55	788.92	51

So, these are the first level of exercise that we do on any of the time series let say I take about 30 gauging locations. Here this is a duration in years which is available there are there is some exception that we make for one of the gauge because of the importance. You may have let say 12 years and more data for all most of the gauges


However, you may have a data length which is quite small; however, because the importance of that particular gauge station, you may have to include that also in the analysis and then you get all the statistics. So, you get the average, you get the maximum flow, minimum flow. This is maximum daily flow and this is a maximum in terms of the discharge here cubic meters per second and this is a volume.

(Refer Slide Time: 50:36)

Data Consistency Checks

Statistical Analysis of Discharge Data:

- The coefficient variation is more than 100% for the gauge sites, Mandla town, Bareli and Tikola, indicating a high variation of flows
- Similar statistics are computed for monthly flow data
- In non-monsoon months a large variation is indicated
- As the flow magnitudes are small in the non-monsoon months, the large variation is not of much practical significance




22

(Refer Slide Time: 50:44)

Statistics of Annual Flows (contd.)

S.No	Gauge site	Data used (Period)	Duration (years)	Average (MCum)	Maximum daily flow (cumec)	Maximum (MCum)	Minimum (MCum)	Standard deviation (MCum)	Coeff. of variation (%)
16	Gadarwara	1977-1999	23	1,304.96	3,080.80	3,016.02	459.23	646.76	50
17	Maheshwar	1985-1993 1996-2000	14	745.17	1,786.00	1,364.76	263.88	299.20	40
18	Bareli	1995-1993 1998-2000	12	756.89	2,964.42	3,749.21	47.10	1,032.01	136
19	Chhidgaon	1976-1999	24	1,007.02	4,460.00	2,187.60	249.97	502.53	50
20	Ginnore	1971-1999	29	2,109.27	12,157.80	4,525.83	553.87	989.50	47
21	Kogaon	1978-1999	22	1,090.14	4,555.00	2,296.16	164.53	660.65	61
22	Ajandiman	1995-1993 1996-2000	14	255.10	1,149.00	625.33	40.74	183.67	72
23	Tikola	1985-1993 1996-1999	13	532.10	1,398.41	1,738.49	61.34	554.75	104
24	Chandwada	1979-1999	21	1,455.58	7,823.80	4,082.33	149.28	1,097.90	75
25	Sandalpur	1987-1993 1996-2000	12	226.10	779.10	423.89	47.77	133.09	59
26	Barmanghat	1988-1999	12	12,842.86	16,283.00	28,749.00	4,052.33	6,120.13	48
27	Balkheri	1977-1999	23	722.18	4,961.10	1,639.75	211.29	350.74	49
28	Barman	1970-1988 1991-1995	24	11,593.41	20,658.20	27,743.94	4,422.63	5,235.99	45
29	Bagratawa	1978-1991	16	1,801.80	9,584.20	5,329.54	61.24	1,383.53	77
30	Gurudeshwar A.M	1970-1976	7	46,756.48	61,000.00	76,885.33	28,104.76	15,774.09	34



31

Like this you get all the statistics also you get the coefficient of variation. Similarly, for all of these locations you get this. Then you go to statistical consistency checks

remember here we have done this exercise after filling up all the missing data. So, once you fill up the missing data, you identify 30 rain gauge thirty stream gauge stations for further analysis and then you start the consistency data.

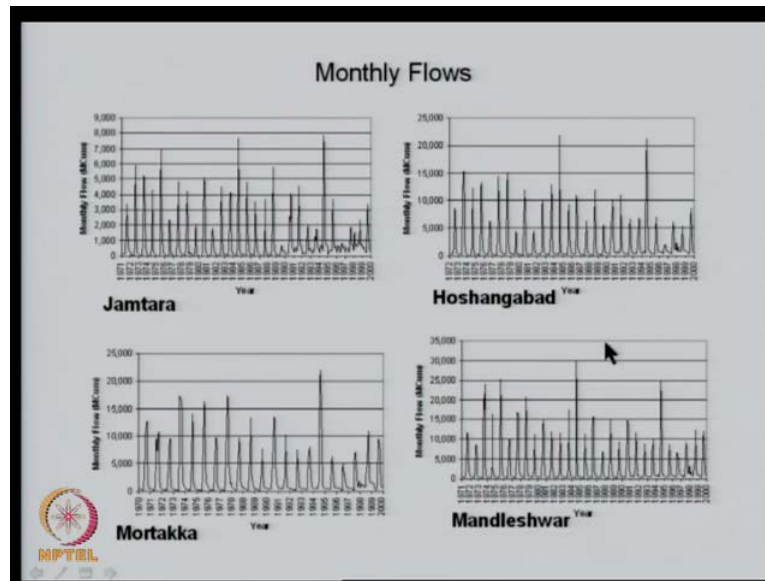
Now, the first level of observation is what level of coefficient of variation exists. So, if you have more than 100 percent you know, it indicates that the variation of flows is quite high for the daily data and similarly, you do with this for monthly as well as annual data.

Specifically for the non monsoon flows, you know which are essentially given by the, which are **which are** contributed by the base flows there will be a large variation at least in this particular example. As the flow magnitudes are themselves small in non monsoon periods for this particular example, the large variation is not of a much practical importance. So, we may be focusing mostly on the monsoon flows.

However, this may not be the case for a Himalayan rivers. Let us say you take Brahmaputra or Ganga river or so on. Now in the non monsoon period also there is a different component of contribution which can be quite significant.

For example glacial melts snow melt and so on and therefore, you must be alert to the situation that the non monsoon flows are contributed not only by the base flow, but, by other mechanisms such as snow melt and glacial melt and therefore, the coefficient of variation will be an important parameter important moment even for the **non** non monsoon flows in such situations.

(Refer Slide Time: 52:40)




Then we plot the time series look at the time series they must be all continuous. So, we have filled all the data and then fill you just plot the time series at various locations. So, these are Jamtara Hoshangabad Mortakka Mandleshwar etc the locations of which are shown in the index diagram that I shown earlier.

(Refer Slide Time: 53:04)

Data Consistency Checks

Specific flows:

- The specific flow is expressed as flow volume per unit area of the catchment
- Represents the catchment response to precipitation
- If a number of gauge stations are located in the same hydroclimatic region with similar land use patterns, then the specific flows computed with data at the gauge stations must be comparable
- Annual specific flows are computed as the ratio of average annual flow to catchment area



24

Then we start computing the specific flows. Like I mentioned the specific flow at a particular gauge will be simply the flow magnitude divided by the area of catchment that is captured by that particular gauge. So, the flow volume per unit area of the catchment.

This represents the catchment responds to precipitation; that means, how is the run off generated for a given precipitation from that particular catchment is given by the specific flow.

So, if you have two gauges one below the other in terms of up being upstream and downstream and if the **if the** catchment is completely homogeneous and if the precipitation is the same; therefore, because of these situations the specific flow that you get in the two gauges must be the same and in practical situations it must be comparable.

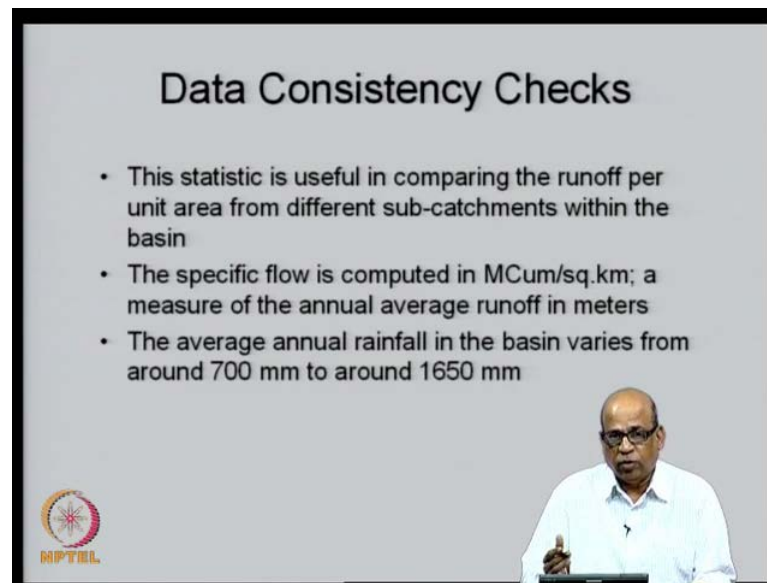
For example if you get a specific flow of .54 units in one of the gauges and you get somewhere around 0.53 or 0.2 these are comparable. However, you get a 0.54 here and 0.17 here or 0.54 here and 0.78 here then these are not comparable. Then we must be look we must be able to look for reasons why it is not comparable.

There may be a physical reasons why they are not comparable your may be your catchments are different may be your precipitation pattern is different or may be your water utilization is different in the two catchments and so on.

So, the fact that the specific flows are different indicates that something is happening in the catchment or something is wrong with our measurement. So, we must be alert to such situation. If your number of rain gauge stations are located in the same hydro climatic region with similar land use patterns then the specific flows computed with the data at the gauge stations must be comparable.

Typically we do this specific flow computations either at annual scale or may be at seasonal scale typically for monsoon season and non monsoon season may not be important for many of the streams in our country. So, at the annual scale as well as at the monsoon scale we may do this specific flow analysis and then look at the consistency.

(Refer Slide Time: 55:40)



Data Consistency Checks

- This statistic is useful in comparing the runoff per unit area from different sub-catchments within the basin
- The specific flow is computed in MCum/sq.km; a measure of the annual average runoff in meters
- The average annual rainfall in the basin varies from around 700 mm to around 1650 mm

MPTEL

So, through this statistic which is a specific flow; we compare the runoff per unit area from different sub catchments within the basin and the specific flow is computed in million cubic meters per square kilometer and a measure of the annual runoff in meters in the depth units.

For this particular basin, the annual rainfall in the basin varies from 700 millimeters to around 1650 millimeters and there is a variation from east to west as you **as you** go from east to west there is a significant variation in the rainfall. Now, all of this will contribute to the change in the specific flows. We will continue this discussion on this particular basin in the next lecture. However, what we did in this particular lecture is going from the mathematical rigor of stochastic hydrology that we have discussed so far in the course.

We now focus on some practical aspects where we start looking at what kind of data we have, what is the quality of the data, what is the consistency of the data with respect to the homogeneity or the lack of homogeneity in the catchment. How do we fill the missing data and construct a continuous time series and so forth.

Now, these are important issues because as I mentioned early on in this lecture, in most of the regions of our country, the data is not of long duration first of all and even where it is available for long duration, there may be missing data and it may be inconsistent with respect to the data that is observed in the same basin by other gauges.

Now, the inconsistency may arise because of several reasons. Let us say that you have a highly sophisticated way of measuring at a particular location A. But, in a surrounding location you have a traditional and perhaps a ancient method of gauging the stream flows. Then the two gauges may give inconsistent readings. But, how do we identify that these two are inconsistent. That is the focus of today's lecture as well as the next lecture.

So, we must be able to identify inconsistency and then correct for the inconsistency. To correct the inconsistency we must know the reason why the inconsistency has occurred. Now, this is a focus of the discussion that we will carry forward in the next lecture.

Thank you for your attention. We will meet again.