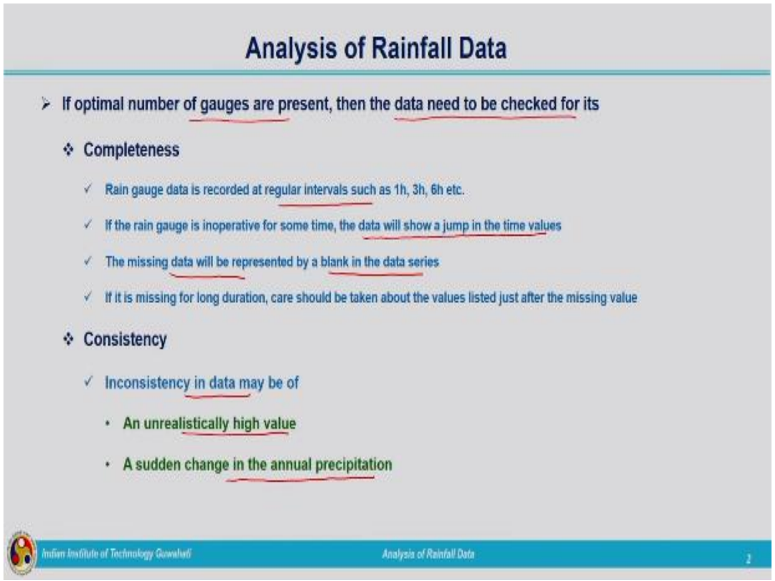
Engineering Hydrology
Professor Doctor Sreeja Pekkat
Department of Civil Engineering
Indian Institute of Technology, Guwahati
Module: 2 - Lecture 21
Analysis of Rainfall Data

Hello all, welcome back. In the previous lecture we were discussing about the representation of rainfall data. We have seen different ways of presenting rainfall data that is by means of histogram, hyetograph and mass curve. In one we were plotting the rainfall depth versus time, second one, we were plotting the intensity versus time and in the case of mass curve, we were plotting the cumulative rainfall versus time.

So, once these plots are ready and the data is there with us that is long series, long years data we are having, so we need to do some hydrologic analysis. In that case whether we will be able to use that data blindly or is there any need of checking the data. So, we need to make sure that the data which has collected is containing accurate data. So, for that we need to have a check of these data before making use of those data for the hydrologic analysis.

(Refer Slide Time: 2:06)



Let us see the analysis related to rainfall data in this lecture. We have seen network of rain gauges and optimal number of rain gauges and if optimal number of rain gauges are present the data which we are using for hydrologic analysis or any purpose need to be checked for its completeness and consistency. Completeness means whether the data series is complete, if 30 years data we are collecting, complete 30 years data is available with us and the data which we have collected is consistent that is there is no inconsistencies related to data. This check needs to be carried out before making use of those data for any of the analysis.

So, what is meant by completeness? Rain gauge collects rainfall at regular intervals, it may be on daily basis, it may be hourly basis or certain rain gauges will be giving you data with 15 minutes, 5 minutes intervals. So, we should get the data with these stipulated intervals.

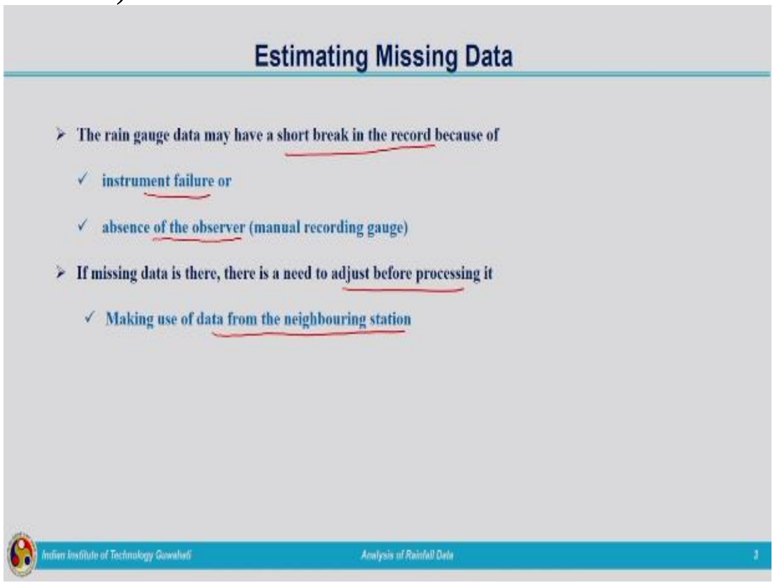
Sometimes it may happen like that the rain gauge may be inoperative. In such cases a large number of data between certain interval of time will be missing. Data will show a jump in the time values. If the rainfall data is recorded at 9 am in the morning because of some issues related to rain gauge it could not record the data for certain time and started recording after 11 o'clock, so in between two hours data related to rainfall is missing. So, this check we need to make sure before making use of these data for any of the applications.

So, the missing data will be usually represented by the blank spaces in the data series and if the missing data is there for long duration maybe for months, days in that way if the missing data is there in the data series then again one more check we need to do regarding the data which we attain just after the missing period, whether the data is corresponding to the previous time duration that is for one hour or one day or it is an erroneous data this check needs to be carried out.

Second one is the consistency check. Inconsistencies in data can occur in certain cases, sometimes it will be showing unrealistically high value. We are having the rainfall series, data series for long period there is no blank spaces but still some data points may be containing errors, showing some very unrealistic high values and sometimes a sudden change in the annual precipitation.

We are expecting an annual precipitation of certain value, beyond that there is a sudden jump, increase or decrease in certain years which was not experienced by us or which was not recorded, then we need to have a check related to consistency of the rainfall data series.

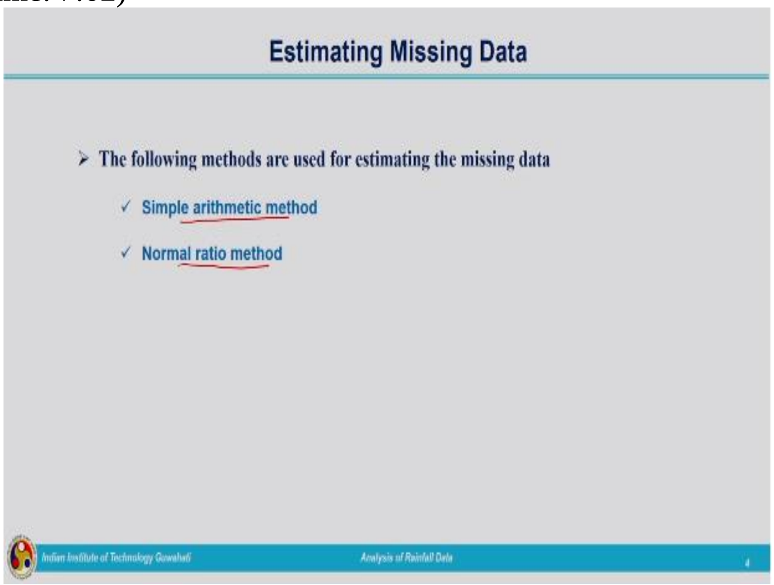
(Refer Slide Time: 5:40)



For estimating missing data, the rain gauge data may have some short break in the record. How can that break occur? It is because instrument failure can happen or absence of the observer in the case of non-recording gauges, non-recording gauges manually we need to observe, so certain days that recording of the rainfall data is missed by the observer and in the case of recording gauges sometimes it may be due to the instrument failure, it can happen or sometimes it may happen that due to heavy rainfall or flooded condition the data recording was not functioning properly and we need to process it before using it for any of the application process.

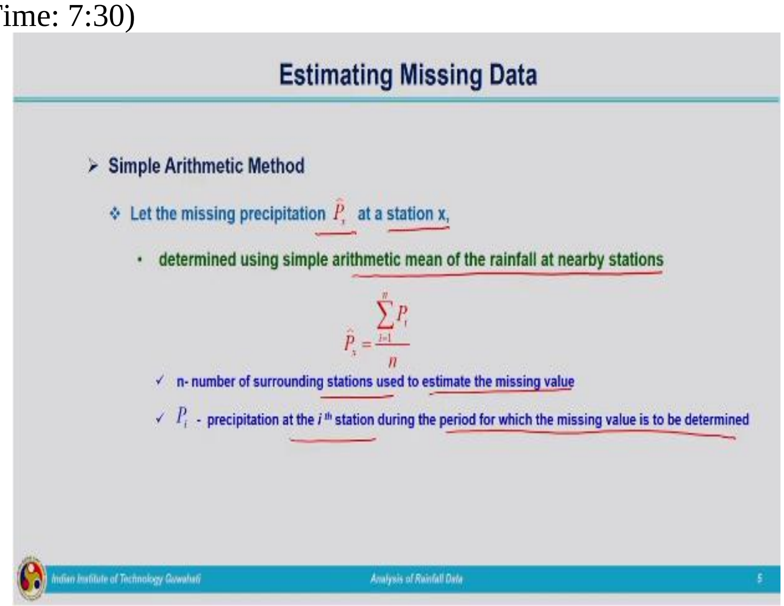
So, what we will be doing, we will be making use of the data from the neighboring stations. Spatially if we check we will be having a network of rain gauges. From the network of rain gauges if one particular rain gauge or the data from a particular rain gauge is consisting of errors or inconsistent or having some missing data what we will be doing, we will be checking it with the neighboring data. The average value corresponding to the neighboring data can be assumed as the missing data.

(Refer Slide Time: 7:02)



Different methods used for estimating missing data are simple arithmetic method and normal ratio method. Simple arithmetic from the name itself it is very clear, just we are averaging out the data from the neighboring stations and second one is a typical method which is only used for this rainfall analysis normal ratio method.

(Refer Slide Time: 7:30)



Let us see the simple arithmetic method first. In this what we are doing, we will be assuming the missing precipitation to be $\hat{P}_x$ at the station *x*, precipitation at station *x* where the data points are missing and it is determined using simple arithmetic mean of the rainfall at nearby stations.

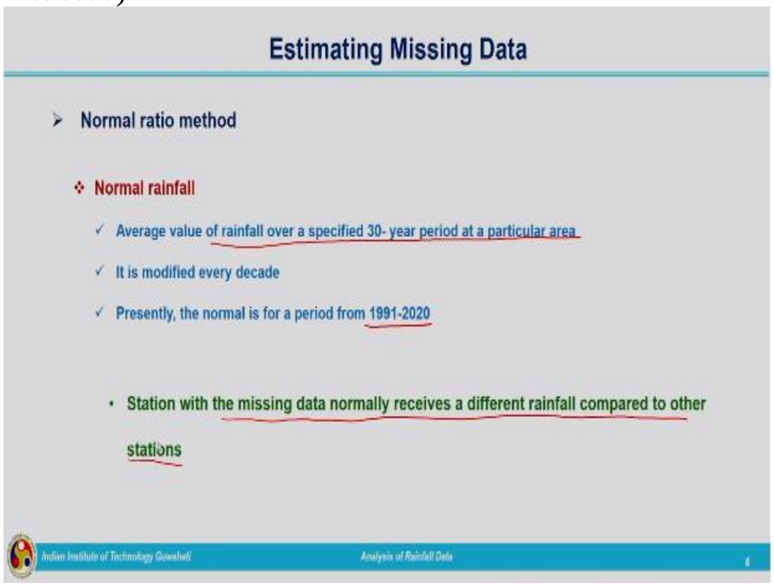
We will be collecting the data from the nearby stations and simple arithmetic average we will be doing and that will be considered as the missing data at this particular station *x*. That can be done by using this formula:

$$\hat{P}_x = \frac{\sum_{i=1}^{n} P_i}{n}$$

In this n is the number of surrounding stations used for estimating the missing value and P_i is the rainfall at i^{th} station during the period for which the missing value need to be determined.

We know certain period for one month or sometimes when we collect time series data for long years, sometimes one-year data itself will be missing. So, in such cases certain other techniques also should be used, so this is the simple method which is used if we are having annual data missing, so we will be using the arithmetic average of the data collected from the neighboring stations that is assumed as the missing data.

(Refer Slide Time: 9:00)



Now, second method is the normal ratio method. Normal ratio method what we are doing? We need to have an understanding about normal rainfall. Normal rainfall is a general term which we are using in hydrology, we need to have an idea what is mean by normal rainfall. Normal rainfall is the average value of rainfall over a specified 30-year period at a particular station that is normal rainfall is calculated from the 30 years data.

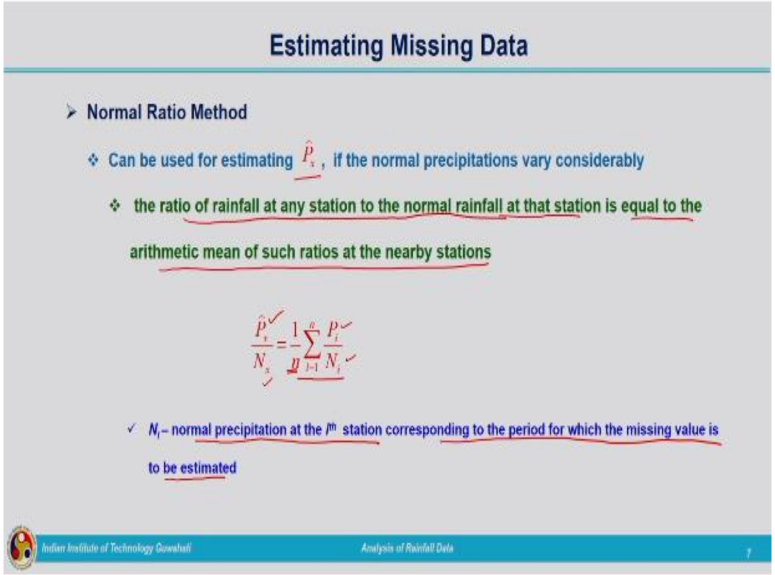
We will be collecting 30 years data that will be average to get the normal rainfall at that particular rain gauge station and this will be modified every decade, that is if we are collecting the data and we have last 120 years data is there, so we can calculate the normal range for 30-year period 1900 to 1930, 31 to 60 that way 91 to 2020.

So, 30, 30 years data we will be considering as the normal rainfall but this rainfall need to be modified every decade. Presently the normal rainfall period is from 1991 to 2020. So, we are in the year of 2021, so we will be making use of the data from 1991 to 2020 and that data averaged out to get the normal data. When it is in the year 2030 we will be using from the year 2001 to 2030 that way every decade we need to modify this data.

Now, what we will be doing in the case of estimating missing data. Station with the missing data normally receives a different rainfall compared to other stations. If there is any missing data for a year or two years or certain period if you are calculating the 30-

year average data there will be deviation from the normal data from the other stations which we are expecting to be almost same.

(Refer Slide Time: 11:13)



So, in this method again we are assuming the missing data as $\hat{P}_x$ and we will be calculating the normal precipitation for that particular station. Then what we will be doing? The ratio of rainfall at any station to the normal rainfall at that station is equal to the arithmetic mean of such ratios at the nearby stations.

One particular station is there in which missing data is recorded, so many blank spaces are there, so then what we will be doing, normal rainfall data for that particular station will be calculated based on the available data and normal rainfall data from the neighboring stations also will be calculated.

After that what we will be doing we will be calculating the missing data $\hat{P}_x$ by using this formula

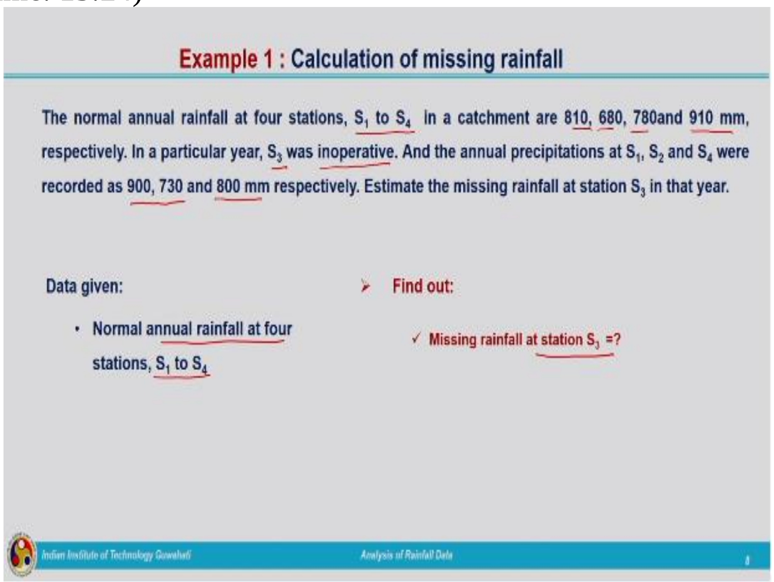
$$\frac{\hat{P}_x}{N_x} = \frac{1}{n}\sum_{i=1}^{n}\frac{P_i}{N_i}$$

Let us see different terms one by one. In this N_i is the normal precipitation at the i^{th} station corresponding to the period for which the missing value is to be determined.

Period should be the same, that is in a particular station we are having certain missing data, so same period we will be checking for the neighboring stations also, $\hat{P}_x$ is the one which we need to determine, N_x is the normal rainfall from that particular station x and that is equated to the average of the ratios we will be calculating for different, different neighboring stations with the rainfall data and the normal rainfall.

Summation of those ratios will be taken and these two ratios will be equated to calculate the missing data at a particular station x and in this you need to n is the number of neighboring stations. How many neighboring stations are considered that number is represented by small letter n, capital N is representing the normal rainfall.

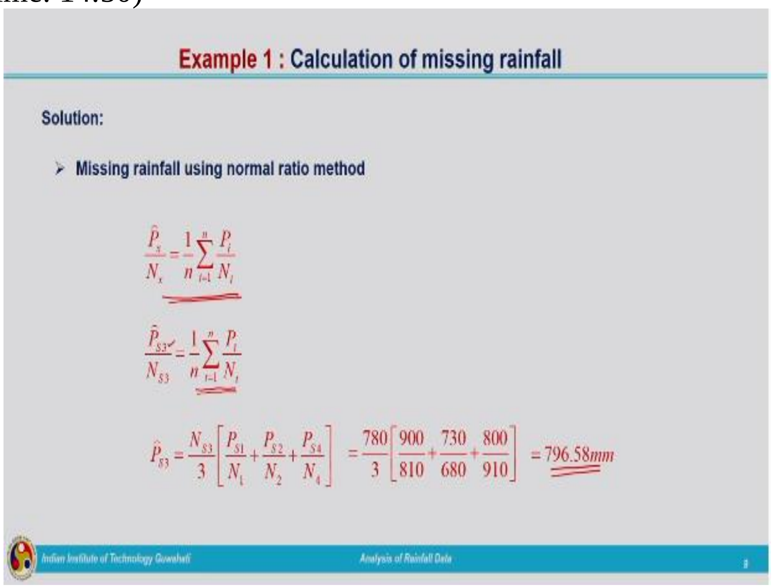
(Refer Slide Time: 13:24)



Now, let us solve one numerical example, then it will be clearer to you. That is the question let me read out first. The normal annual rainfall at four stations S_1 to S_4 in a catchment are 810, 680, 780 and 910 millimeters respectively. In a particular year S_3 was inoperative and the annual precipitation at S_1, S_2 and S_4 were recorded as 900, 730 and 800 millimeters respectively. Estimate the missing rainfall data at station S_3 in that particular year.

So, which are the data given data, normal annual rainfall at four stations S_1 to S_4 are already given to us, one-year data, one-year rainfall data is missing at the station 3, so we need to calculate the annual rainfall corresponding to that particular station for that missing period. Missing rainfall data at the station S_3 need to be calculated, that is the annual rainfall data for that particular year we need to calculate based on the data from the neighboring station and also the normal rainfall at this particular station.

(Refer Slide Time: 14:50)



We will be just making use of the formula which we have seen and after substituting the given data we can get the annual rainfall at the station x. Missing rainfall using normal ratio method we need to use, so the formula we have already seen,

$$\frac{\hat{P}_x}{N_x} = \frac{1}{n}\sum_{i=1}^{n}\frac{P_i}{N_i}$$

We are finding out the ratios of the annual rainfall and the normal rainfall at each and every station, then average those values and multiply it with the normal rainfall of the station x.

So, here we are just substituting the values, our station x is the third station S_3,

$$\frac{\hat{P}_{S3}}{N_{S3}} = \frac{1}{n}\sum_{i=1}^{n}\frac{P_i}{N_i}$$

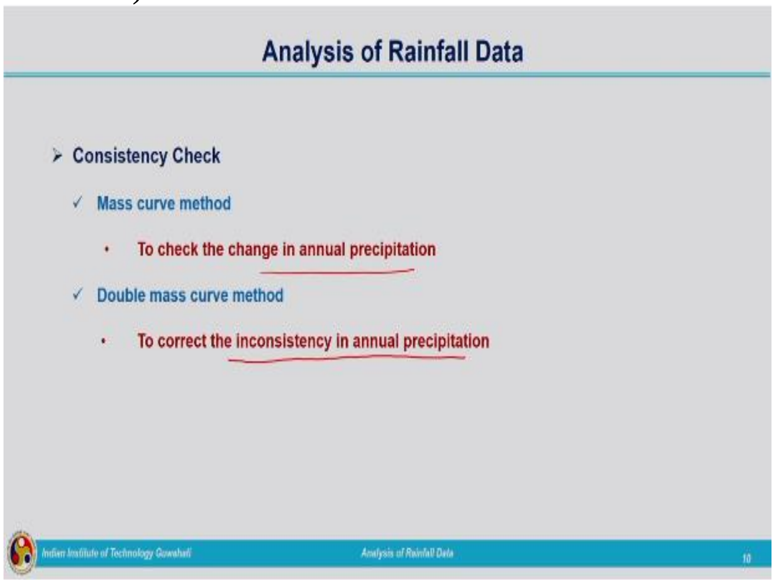
Just we need to substitute the values that is P_{S3} can be calculated by substituting the given values and

$$\hat{P}_{S3} = \frac{N_{S3}}{3}\left[\frac{P_{S1}}{N_1} + \frac{P_{S2}}{N_2} + \frac{P_{S4}}{N_4}\right]$$

$$= \frac{780}{3}\left[\frac{900}{810} + \frac{730}{680} + \frac{800}{910}\right] = 796.58mm$$

So, this is the way in which the missing data can be calculated for a particular station by making use of the annual rainfall data and the normal rainfall data from the neighboring stations.

(Refer Slide Time: 16:02)

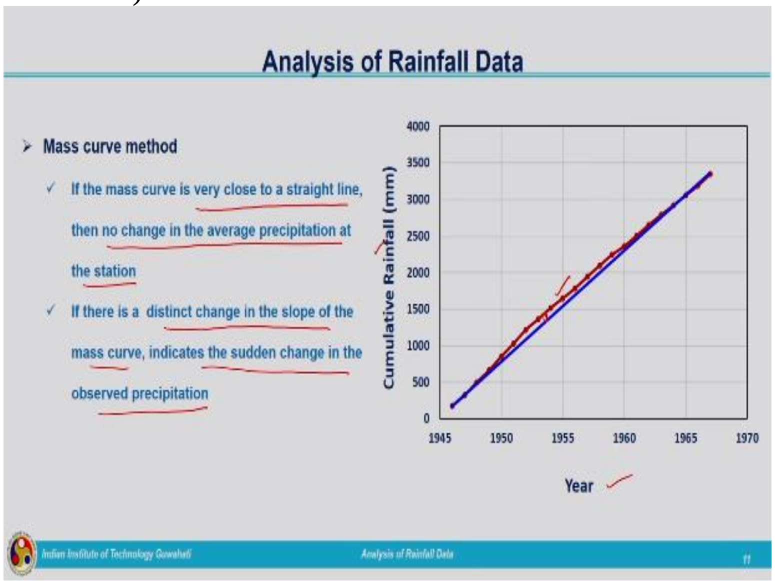


Now, second check is regarding the consistency check. Two checks we need to do, that is the missing value check and also the consistency check. We need to check whether the data which we have collected is consistent or not. For consistency check we will be making use of mass curve method. Mass curve method we know what is meant by mass curve.

For a particular station we will be calculating the cumulative rainfall and cumulative rainfall versus time will be plotted, that is the mass curve. Rainfall data collected at a particular station, incremental rainfall data will be there with you or depending on the type of data collected we need to get the cumulative data for each and every interval, that will be plotted against time to get the mass curve.

So, here for consistency check first method is to make use of the mass curve method that is if you are plotting the mass curve we can check the change in annual precipitation. Then second method is double mass curve method this is used for making the correction in the annual precipitation. The inconsistency which is present in the rainfall data series can be corrected by making use of double mass curve method.

(Refer Slide Time: 17:29)

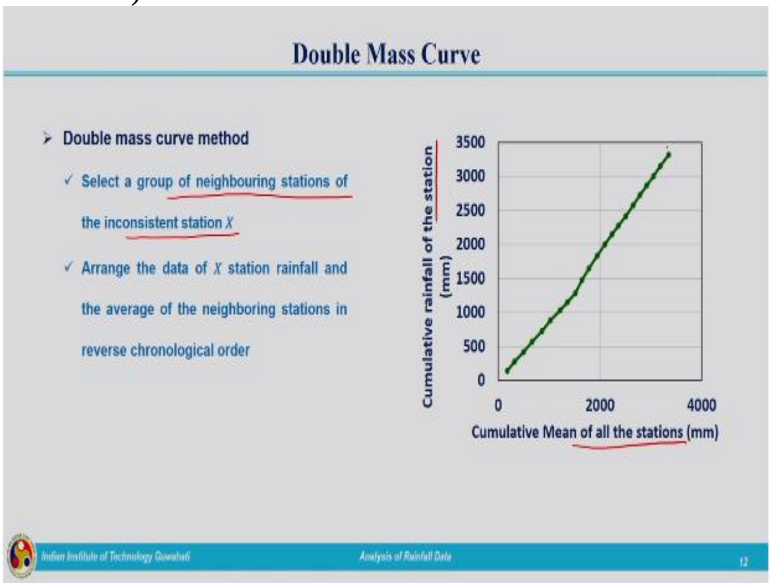


Let us see one by one. In mass curve method what we are doing, mass curve how it is plotted we have seen in the previous lecture. If the mass curve is very close to a straight line then no change in the average precipitation at the station. So, first we will be plotting the mass curve and we will check the curve, if it is very close to a straight line then we can assume that there is no inconsistency. This way if we have plotted the cumulative rainfall versus time then we are making a straight line and see whether there is any variation from the straight line.

So, if there is a distinct change in the slope of the mass curve that indicates a sudden change in the observed precipitation. So, mass curve method we are having the data series, we are constructing the mass curve corresponding to the collected data, if there is any sudden change in the slope of the mass curve or we will check with the straight line whether there is any deviation from the straight line corresponding to particular year data, then we need to go for correction of this data.

So, from this we can only observe that whether there is any inconsistency or not, any particular year is showing any inconsistent data or not that can be understood from the mass curve method. Now, if there is any inconsistency, here in this figure you can see there is variation from the straight line, so this line is slightly above the straight line. So how can we correct this deviation? So that can be done by means of double mass curve.

(Refer Slide Time: 19:11)



So, in double mask curve method what we are doing, we are selecting a group of neighboring stations near to the inconsistent station *X* and after that what we will do, arrange the data of station *X*, arrange the data of station *X* and the neighboring stations in the reverse chronological order, that is from 1980 to 2020 we are having data, we will be arranging the data in such a manner that 2020 to 1980 that way reverse chronological order we will be arranging the data.

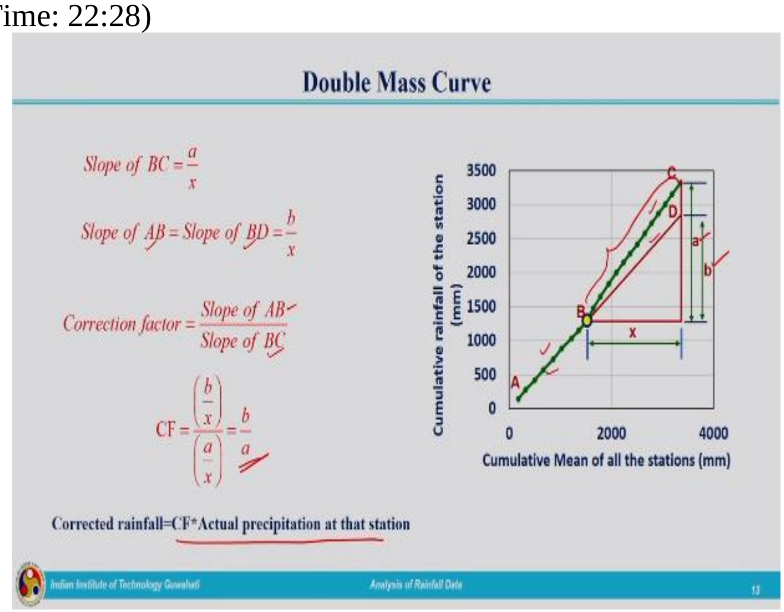
Why we are doing this? We are assuming that there will not be much error in the recent years data because recent years data are mainly from the recording rain gauges. Majority of the non-recording rain gauges have been replaced by the recording rain gauges and functioning well than that of the old rain gauges. So, the old data in the data series may contain some errors that is our assumption. So, based on that we are going to make the correction to make the old data to come in same line with the new data.

So, here first we will be arranging the data in reverse chronological order and next step is to average out the data from the neighboring stations. If six neighboring stations data are there, we will find out the average of rainfall data from all these stations and we are having the data from the particular station which is having inconsistent data. After that what we will do? We need to plot the mass curve for that we have to calculate the cumulative rainfall data. Once the cumulative rainfall data for each and every interval is ready we will plot the double mass curve.

How this double mass curve is plotted? Along the y-axis we are plotting the cumulative rainfall data of the inconsistent station and along the x-axis cumulative mean of all the neighboring stations. So, the data which we have selected along the y-axis will be using the inconsistent data and for along the x-axis we will be using the average data from the neighboring station.

These two data are the cumulative data, we are plotting the ordinates of the mass curve for the inconsistent station and the average of the neighboring stations, cumulative rainfall data will be plotted. So, you can see this green line. Here in this curve, if you are observing any sudden change in the slope that is indicating there is an inconsistency in the data corresponding to those years. If the observable slope change is not there then we can assume that the data is consistent.

(Refer Slide Time: 22:28)



So, here in this figure you can see at this particular point there is a change in slope taking place, so I am naming the three points beginning of the curve A, B, C. This way I am naming and AB is the recent years data, we have arranged the data in reverse chronological order, so our graph will be beginning from the recent years. A to B is the recent years data that we will be extending by means of a straight line, after that what we will be doing we will be finding out the slopes of the recent data and also for the line where we have observed the slope change, slope variation. So, for that we are finding out the ordinates and abscissa.

So, this BC is representing the inconsistent data and we are assuming that AB is correct because when this plotted at a particular point B we have seen there is a change in slope taken place. So, now we are going to find out the slope of these two lines AB and BC. For that we are taking the ordinates and abscissa, so slope of BC can be obtained by taking the ratio *a* divided by *x*.

$$\text{Slope of BC} = \frac{a}{x}$$

Now, coming to slope of BD we need to have the ordinate that that can be represented by *b*. So, slope of AB and slope of BD are same, it is the same line only, BD is the extension of AB. So, slope of AB and slope of BD are same it can be calculated by taking the ratio *b* by *x*.

$$\text{Slope of AB} = \text{Slope of BD} = \frac{b}{x}$$

Now, we need to check the difference in slope how much is the difference in slope coming. If you calculate it beyond 10%, if the value difference in slope is more than 10% there is a need of correction required, if it is within 10% then we can keep the data as such.

Now, in what way we will be doing the correction. We need to find out some correction factor. That correction factor can be calculated by taking the ratio of slope of AB divided by slope of BC.

$$\text{Correction factor} = \frac{\text{Slope of AB}}{\text{Slope of BC}}$$

Slope of AB is *b* by *x* and slope of BC is *a* by *x*, so correction factor will be given by the ratio *b* divided by *a*, i.e.,

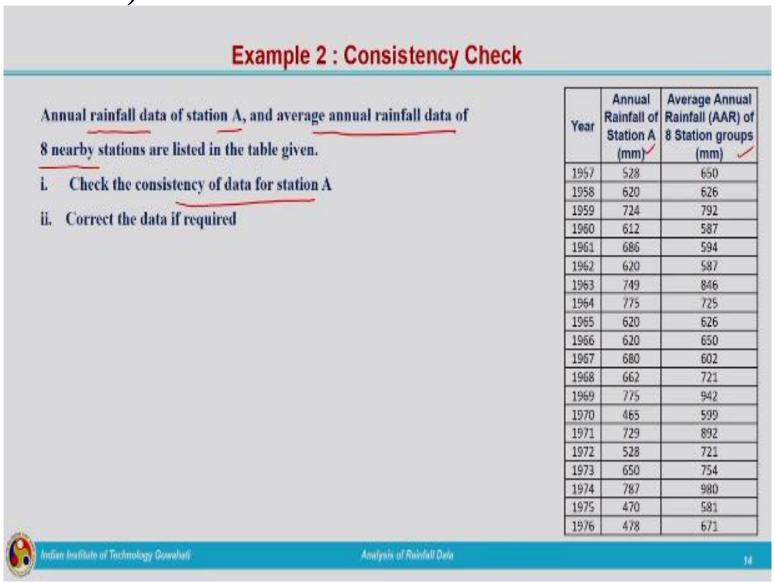
$$CF = \frac{\left(\frac{b}{x}\right)}{\left(\frac{a}{x}\right)} = \frac{b}{a}$$

Only we need to get the ordinance *a* and *b*, if *a* and *b* are known to us we can get the correction factor.

After that what we will do we will multiply the data corresponding to this BC line, this particular line will be multiplied with this correction factor to make it align with the new series, that is what we will be doing in the case of double mass curve. We are correcting the past data based on the new data. So, this will be clearer to you when we solve a numerical example and corrected rainfall is correction factor multiplied by actual precipitation at that station, i.e.,

$$\text{Corrected Rainfall} = \text{CF} \times \text{ Actual precipitation at that station}$$

(Refer slide Time: 25:44)

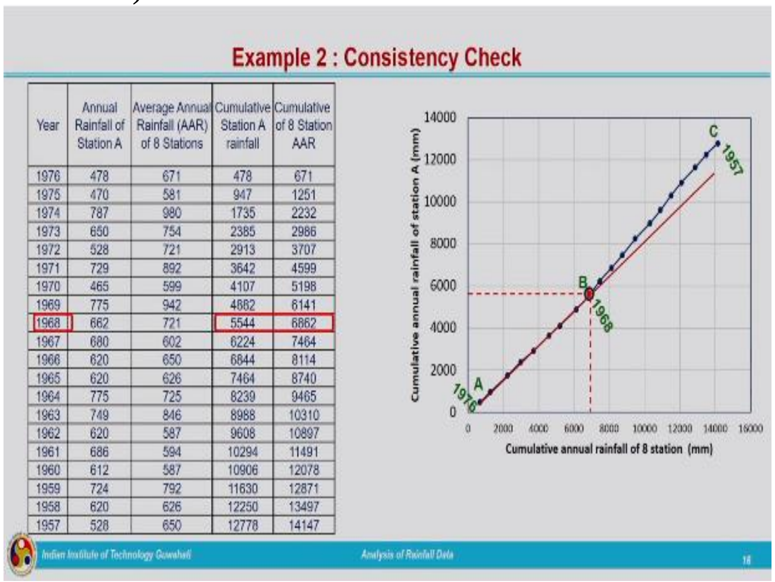


Year	Annual Rainfall of Station A (mm)	Average Annual Rainfall (AAR) of 8 Station groups (mm)
1957	528	650
1958	620	626
1959	724	792
1960	612	587
1961	686	594
1962	620	587
1963	749	846
1964	775	725
1965	620	626
1966	620	650
1967	680	602
1968	662	721
1969	775	942
1970	465	599
1971	729	892
1972	528	721
1973	650	754
1974	787	980
1975	470	581
1976	478	671

Now, we can solve a numerical example related to consistency check. Let me read out the question. Annual rainfall data at station A and average rainfall data of 8 nearby stations are listed in the table given below. So, we are having the annual rainfall data of station A and average annual rainfall data of 8 nearby stations. So, it is given in the table, we are having the annual rainfall data at station A and also annual rainfall data from the 8 stations were given and average of that has been taken and put in the table.

Now, what we need to do? We need to check the consistency of data for station A, whether this data series is consistent or not and second one is if there is any inconsistency you need to correct it by making use of double mass curve method. If you are using mass curve method you can understand whether all the data points are coming on a straight line then or there is no deviation in slope, observed deviation is not there then we can assume that the data is consistent or for doing the correction you can go for finding out the correction factor by means of the double mass curve. Here this data I am not doing the mass curve approach because I have already checked and I know it is having some errors. So, we will solve the problem by using double mass curve.

(Refer Slide Time: 27:16)



Year	Annual Rainfall of Station A	Average Annual Rainfall (AAR) of 8 Stations	Cumulative Station A rainfall	Cumulative of 8 Station AAR
1976	478	671	478	671
1975	470	581	947	1251
1974	787	980	1735	2232
1973	650	754	2385	2986
1972	528	721	2913	3707
1971	729	892	3642	4599
1970	465	599	4107	5198
1969	775	942	4882	6141
1968	662	721	5544	6862
1967	680	602	6224	7464
1966	620	650	6844	8114
1965	620	626	7464	8740
1964	775	725	8239	9465
1963	749	846	8988	10310
1962	620	587	9608	10897
1961	686	594	10294	11491
1960	612	587	10906	12078
1959	724	792	11630	12871
1958	620	626	12250	13497
1957	528	650	12778	14147

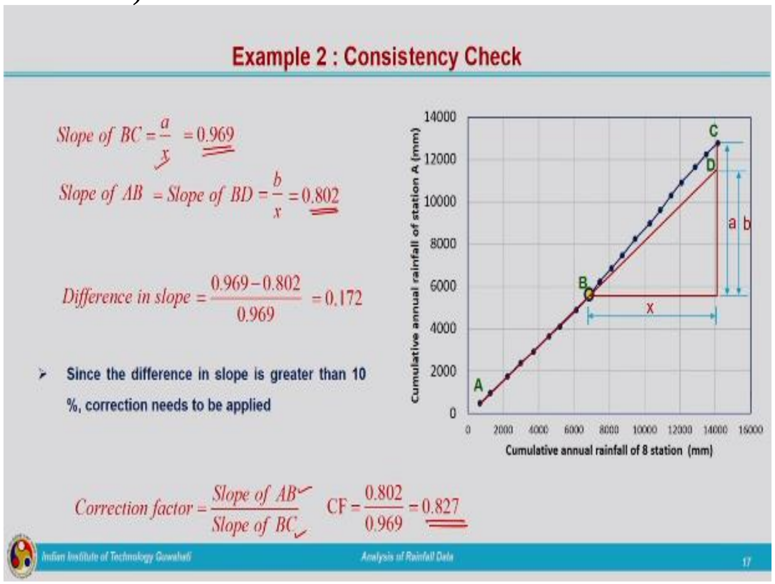
What we have done first. We have done the arrangement of data in the reverse chronological order, annual rainfall data of station A and also from the average data from 8 neighboring stations we have arranged. After that we need to plot the double mass curve. We are plotting the cumulative rainfall at station A versus cumulative rainfall, average of these rainfall from the neighboring stations. So, we need to calculate the cumulative rainfall corresponding to station A and also corresponding to cumulative of 8 neighboring stations.

So, then we will be plotting the double mass curve, along the y-axis we are having the cumulative rainfall from the annual rainfall from station A and along the x-axis we are plotting from the neighboring 8 stations. Units can be in centimeters or millimeters depending on your convenience. Now, we will look at the graph whether there is any observable change in the slope.

So, you can see this is again I am marking the year in order to make it clear, it is starting from 1976 to 1957 and I am fitting a curve straight line. You can see some of the data is deviating from the straight line. So that point you can mark it by this yellow dot. From this point onwards, the data deviation is starting.

So, now we need to understand the data corresponding to which year. So those data points have been checked, it is coming between 6000 to 7000 and 5000 to 6000. So that data can be identified from the table it is corresponding to the year 1968. This data that is from where the change in slope is taking place that is from the year 1968. So, we need to correct the data before 1968. We are assuming that from 1976 to 1968 the data is correct and inconsistency started from the year 1968. So, after finding out the correction factor we will be correcting the data corresponding to the years 1968 to 1957. Let us name this point A, B, C because we need to calculate the slopes corresponding to these two lines.

(Refer Slide Time: 29:47)



We need to find out the ordinance *a*, *b* and *x*. So, slope of BC is the line representing the inconsistent data that is given by

$$\text{Slope of BC} = \frac{a}{x}$$

So, from the table or from the graph we can get the values and it was calculated to be 0.969 and in the similar way we need to calculate the slope of BD or slope of AB, slope of AB is same as slope of BD. Slope of AB and BD are same that can be calculated by taking the ratio *b* by *x* it is calculated

$$\text{Slope of AB} = \text{Slope of BD} = \frac{b}{x} = 0.802$$

Now, what we need to see, we need to check whether the difference in slopes are within 10% or more than 10%.

$$\text{Difference in slope of AB} = \frac{0.969 - 0.802}{0.969} = 0.172$$

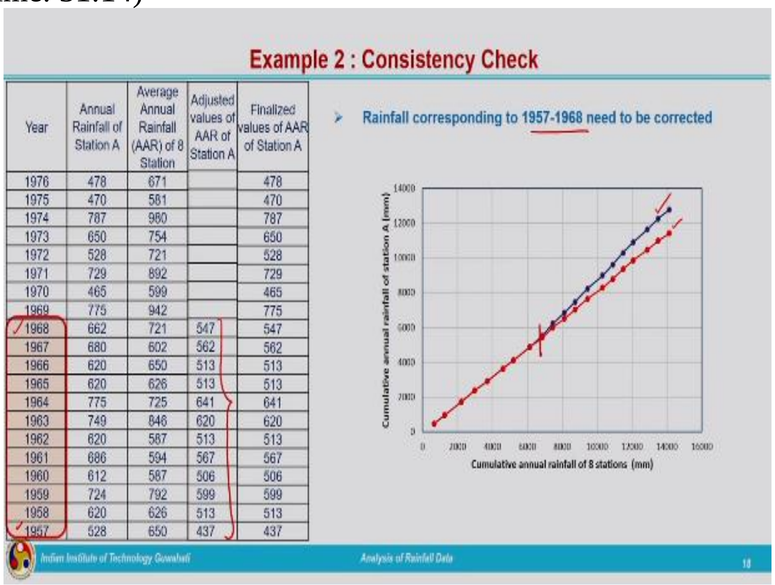
Difference in slope is calculated to be 0.172 that is around 17%. If it is more than 10% we need to go for correction, we need to apply the correction. So, next step is to calculate the correction factor. Correction factor can be calculated by finding the ratio of these slopes,

$$\text{Correction factor} = \frac{\text{Slope of AB}}{\text{Slope of BC}}$$

$$CF = \frac{0.802}{0.969} = 0.827$$

We are already having the corresponding slopes and the value is calculated to be 0.827. Now, we need to apply this correction for the data from 1957 to 1968.

(Refer Slide Time: 31:14)



Year	Annual Rainfall of Station A	Average Annual Rainfall (AAR) of 8 Station	Adjusted values of AAR of Station A	Finalized values of AAR of Station A
1976	478	671		478
1975	470	581		470
1974	787	980		787
1973	650	754		650
1972	528	721		528
1971	729	892		729
1970	465	599		465
1969	775	942		775
1968	662	721	547	547
1967	680	602	562	562
1966	620	650	513	513
1965	620	626	513	513
1964	775	725	641	641
1963	749	846	620	620
1962	620	587	513	513
1961	686	594	567	567
1960	612	587	506	506
1959	724	792	599	599
1958	620	626	513	513
1957	528	650	437	437

We will go back to our data again. So, rainfall corresponding to 57 to 68 need to be corrected. So, from this 57 to 68 need to be corrected. So those data series we will be checking in multiplying it with the correction factor.

So, after multiplying with the correction factor we will be obtaining the rainfall data like this and the entire series will be that is 1968 to 76, recent data will be maintained as such and before that what is there which was showing the inconsistency were multiplied with the correction factor.

Now, we can plot and see whether it is coming in the straight line or not. So, you can see with the inconsistent data points we were having the blue curve and after the correction we have checked that is coming in the same straight line without showing any change in slope.

So, this way we can correct that, that is we were having the data up to this point from 1976 to 69, we were assuming that it is not having any inconsistency and 57 to 68 we have found there is some inconsistency. We are not going after what is the reason behind it, we need to correct it before using it for any of the hydrological applications. So that can be done by using the double mass curve. So, this much about the consistency check.

So, in today's lecture what we have seen. We have seen how the data, rainfall data can be analyzed, so this analysis need to be carried out for checking the missing data and also for checking the consistency in case any inconsistent data is present in the data series. So, we need to replace the missing data by making use of the methods which I have explained and also, we need to check the inconsistency if there are any corrections need to be applied by making use of the correction factor with the help of double mass curve.

(Refer Slide Time: 33:28)

So, here I am winding up this lecture. So more numerical examples can be obtained in these reference textbooks. You have to go through these textbooks in addition to the video lecture. Thank you.