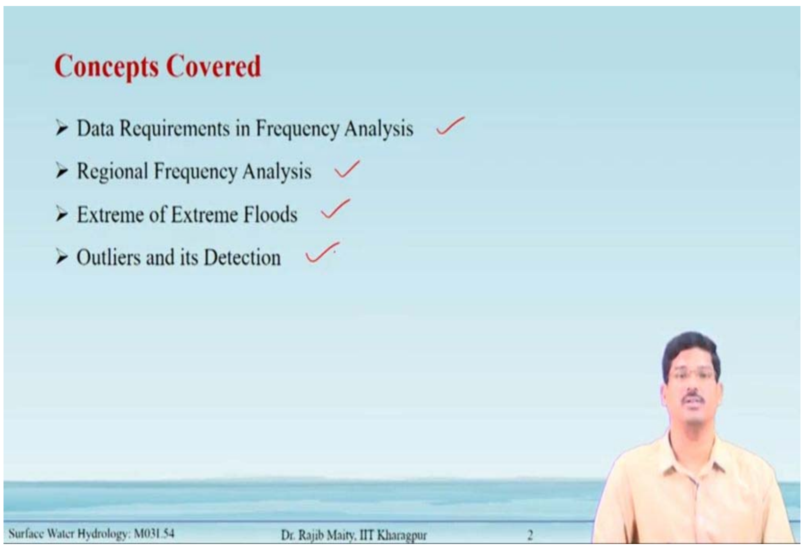
Surface Water Hydrology
Professor Rajib Maity
Department of Civil Engineering
Indian Institute of Technology, Kharagpur
Lecture – 54
Various Issues behind Frequency Analysis

In this lecture, we will discuss some of the issues related to the data that we should consider, while we are going for this frequency analysis, and we should keep in mind the quality of this data.

(Refer Slide Time: 00:40)

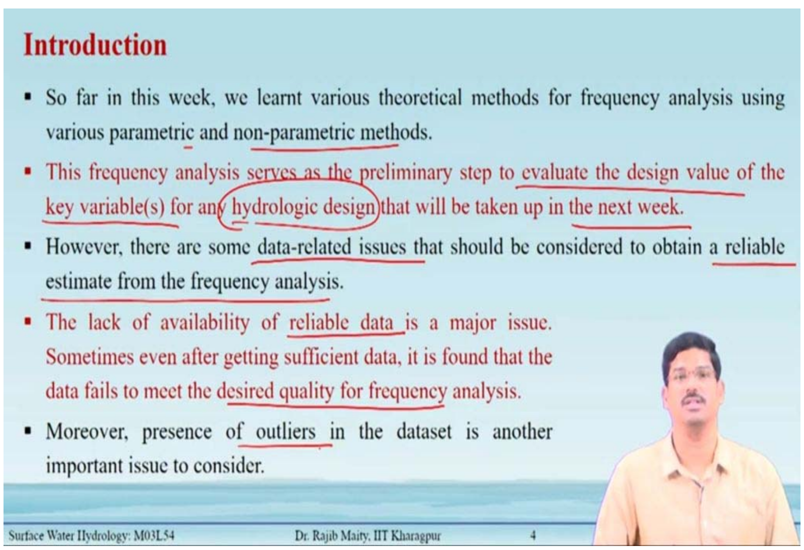


So, there are different issues that we will discuss that are written under this concept cover. The first one is the data required in the frequency analysis, what are the different issues when going for the regional frequency analysis for over a region, and what are the things that have to be considered. And then there are some extreme floods all over the world. And last but not least is the outliers and their detection, which is very sensitive when we are seeing that so many extreme values are coming in due to this climate change impact.

(Refer Slide Time: 01:18)

The outline for this lecture first will give a brief introduction to why these issues are important. Then we will take those issues one after another as I told under this concept cover, and we will also take one example problem before we go to the summary.

(Refer Slide Time: 01:37)

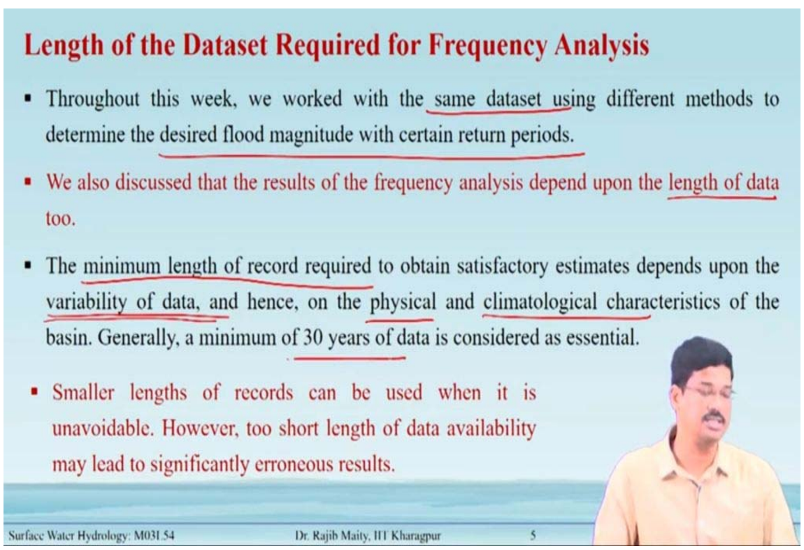


Introduction

This week, we learned the various theoretical method for frequency analysis using the various parametric and non-parametric methods. And this frequency analysis serves as a preliminary step to evaluate the design value of the key variables of the hydrologic design. So, the frequency analysis helps us to evaluate the design value for those key variables. However, that is why there are some data-related issues that we need to consider before applying all these methods that we have learned to get a reliable, reliable estimate from the frequency analysis.

The lack of availability of reliable data is the first major issue, and sometimes even after getting sufficient data, it is found that data fails to meet the desired quality for the frequency analysis. Now, what is that desired quality that we should check and second thing is that the presence of outliers in this data set is an important issue?

(Refer Slide Time: 03:27)

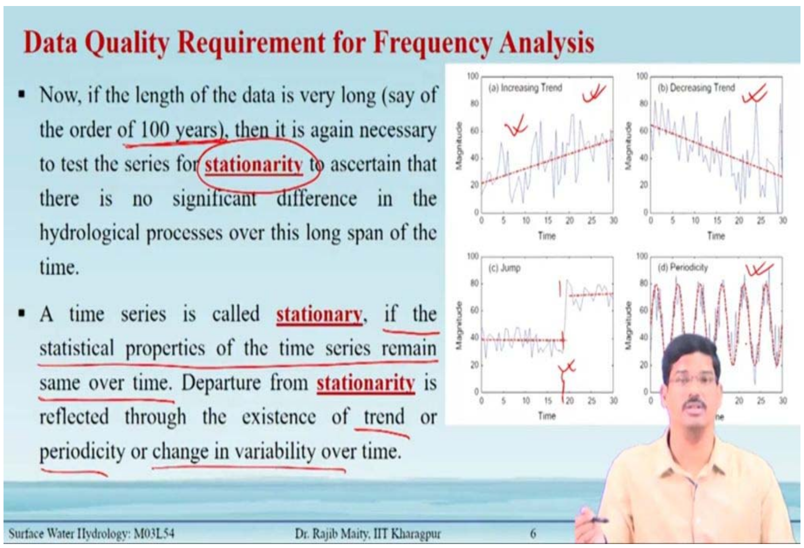


Length of the Dataset Required for Frequency Analysis

The minimum length that is required to obtain a satisfactory estimate depends upon the variability of this data. So, that is one of the important factors that decide the should be a minimum length of the recorded data to get yes, the reliable properties of the data set that can be extracted now. So, that means the variability of the data might have been controlled by different factors like the physical or the climatological characteristics of the basin. Generally, we say that if we get 30 years of this data that can be considered as the essential longer is of course, very much useful.

When you go for the smaller length of this record that also can be used if it is sometimes not unavoidable, we have to remember that there may be some parameters that need to be updated. However, if the data length is too short due to the lack of data availability that may, may lead to a significantly erroneous estimate we have to keep in mind.

(Refer Slide Time: 05:17)

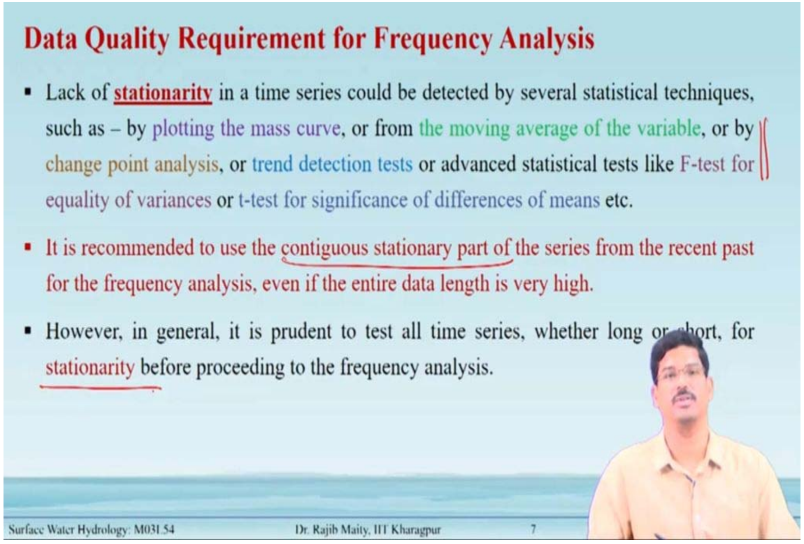


Data Quality Requirement for Frequency Analysis

Now, if the length of the data is very long (say of the order of 100 years), then it is again necessary to test the series for stationarity to ascertain that there is no significant difference in the hydrological processes over this long span of the time, which is a very difficult and tricky question because we know that the processes may change or the impact from the external forces like climate change and all can change the temporal structure of any time series at any location.

A time series is called stationary if the statistical properties of the time series remain the same over time. Departure from stationarity is reflected through the existence of trend or periodicity or change in variability over time.

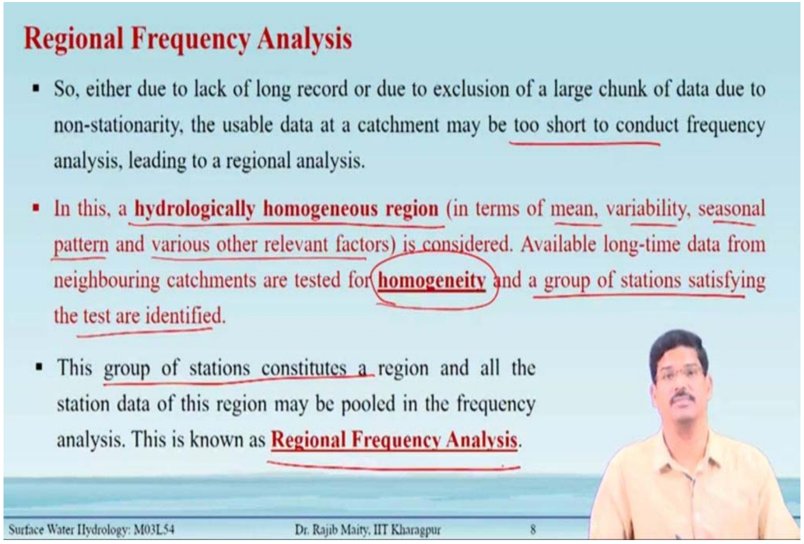
(Refer Slide Time: 07:32)



Lack of stationarity in the time series could be detected by the several statistical techniques are there, such as plotting the mass curve or the moving average of this variable, or by change 0. Analysis, or the trend detection test, or the advanced statistical test like the F-test for the equality of the variance, or the t-test for the significant difference in this mean etcetera. These are some of the tests that utilize this advanced statistical technique. But, of course, is a need for this study to understand how really to detect the stationarity of a time series, and then can be applied.

So, it is also recommended to use the contiguous stationary part of the series, this part is stationary, then that is what we utilize so far as the frequency analysis is concerned. And of course, there are many other applications are there, where it demands to be the data to be first of all, all these trend component and periodicity components should be taken out. However, in general, it is prudent to test all the time series, whether it is long or short, first of all, to test for this stationarity before proceeding to frequency analysis.

(Refer Slide Time: 08:58)



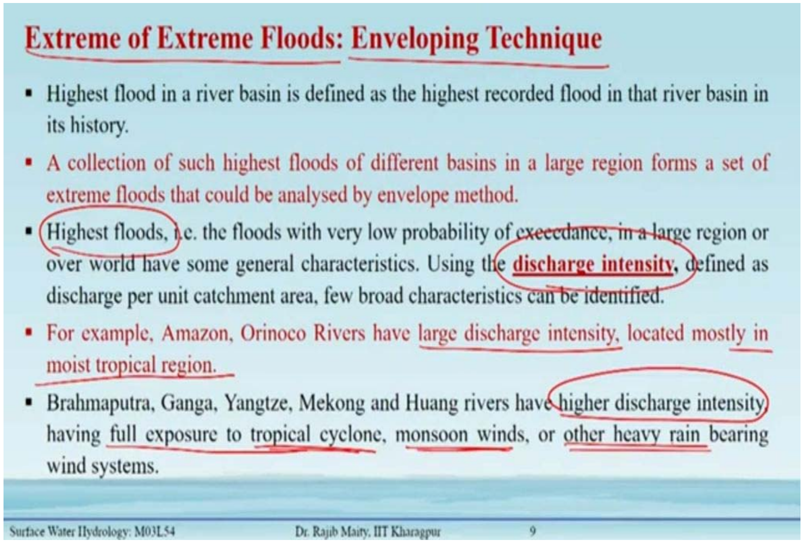
Regional Frequency Analysis

So, either due to a lack of long records or due to exclusion of a large chunk of data due to non-stationarity, the usable data at a catchment may be too short to conduct frequency analysis, leading to regional analysis.

In this, a hydrologically homogeneous region (in terms of mean, variability, seasonal pattern, and various other relevant factors) is considered. Available long-time data from neighboring catchments are tested for homogeneity and a group of stations satisfying the test is identified.

This group of stations constitutes a region and all the station data of this region may be pooled in the frequency analysis. This is known as Regional Frequency Analysis.

(Refer Slide Time: 10:07)

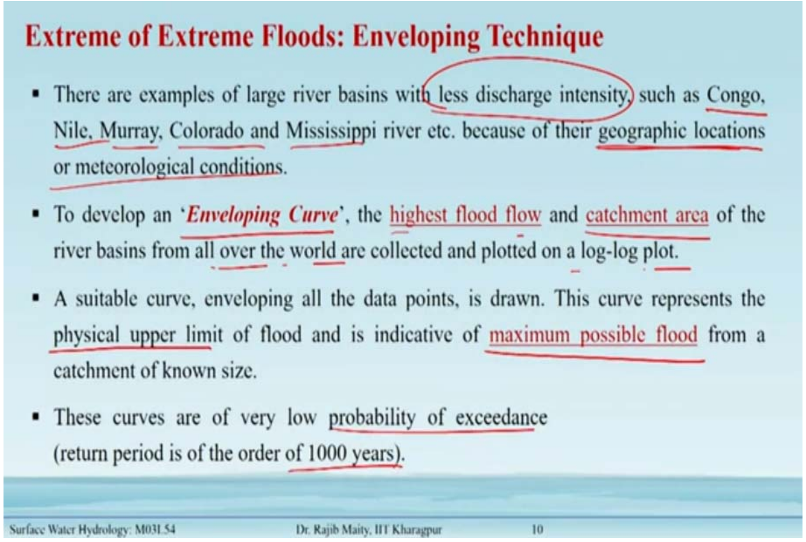


Extreme of Extreme Floods: Enveloping Technique

The highest flood in a river basin is defined as the highest recorded flood in that river basin in its history. A collection of such highest floods of different basins in a large region forms a set of extreme floods that could be analyzed by the envelope method.

The highest floods, i.e., the floods with a very low probability of exceedance, in a large region or over world have some general characteristics. Using the discharge intensity, defined as discharge per unit catchment area, a few broad characteristics can be identified. For example, Amazon, and Orinoco Rivers have large discharge intensity, located mostly in the moist tropical region. Brahmaputra, Ganga, Yangtze, Mekong, and Huang rivers have higher discharge intensity, having full exposure to a tropical cyclone, monsoon winds, or other heavy rain-bearing wind systems.

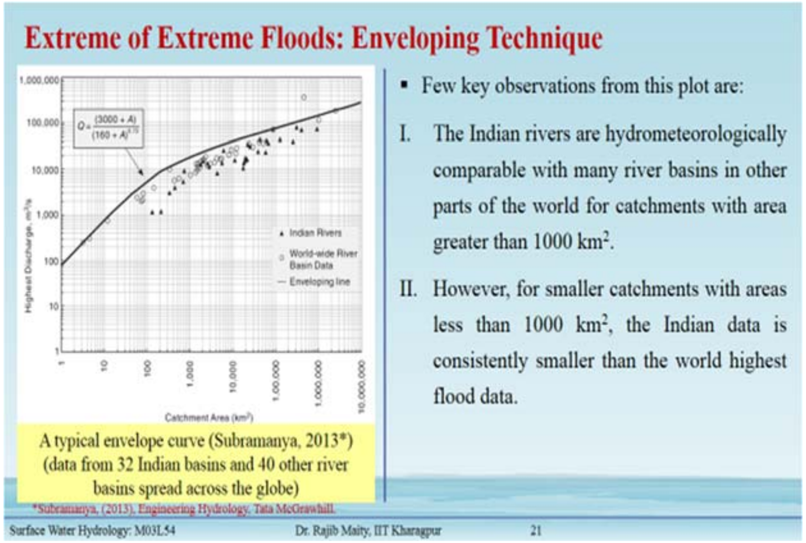
(Refer Slide Time: 11:22)



Similarly, there are examples of large river basins with less discharge intensity, such as Congo, Nile, Murray, Colorado, and Mississippi River, etc. because of their geographic locations or meteorological conditions.

To develop an 'Enveloping Curve', the highest flood flow and catchment area of the river basins from all over the world are collected and plotted on a log-log plot. A suitable curve, enveloping all the data points, is drawn. This curve represents the physical upper limit of flood and is indicative of the maximum possible flood from a catchment of known size. These curves are of very low probability of exceedance (return period is of the order of 1000 years).

(Refer Slide Time: 12:51)



A few key observations from this plot are:

I. The Indian rivers are hydro meteorologically comparable with many river basins in other parts of the world for catchments with an area greater than 1000 km^2.

II. However, for smaller catchments with areas less than 1000 km^2, the Indian data is consistently smaller than the world's highest flood data.

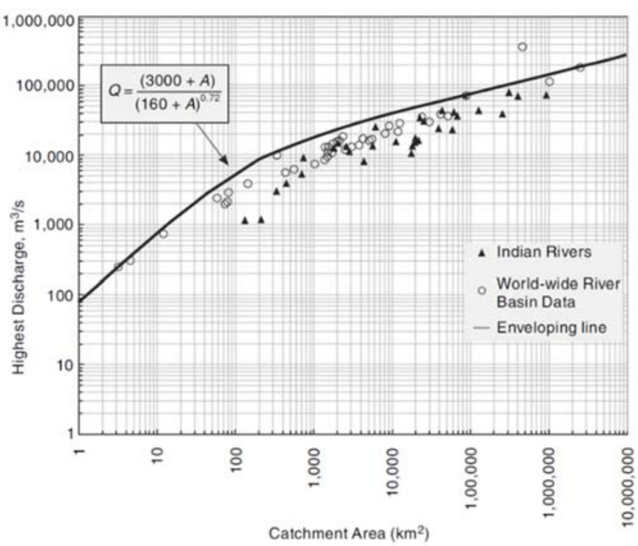
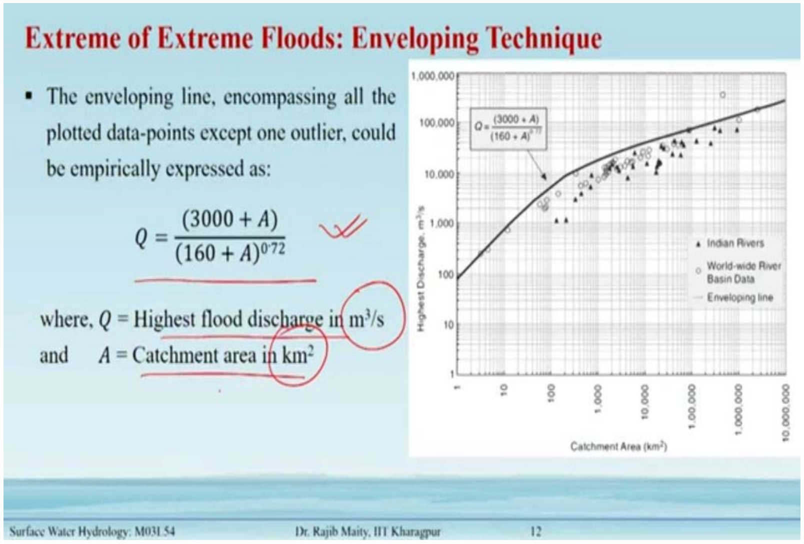
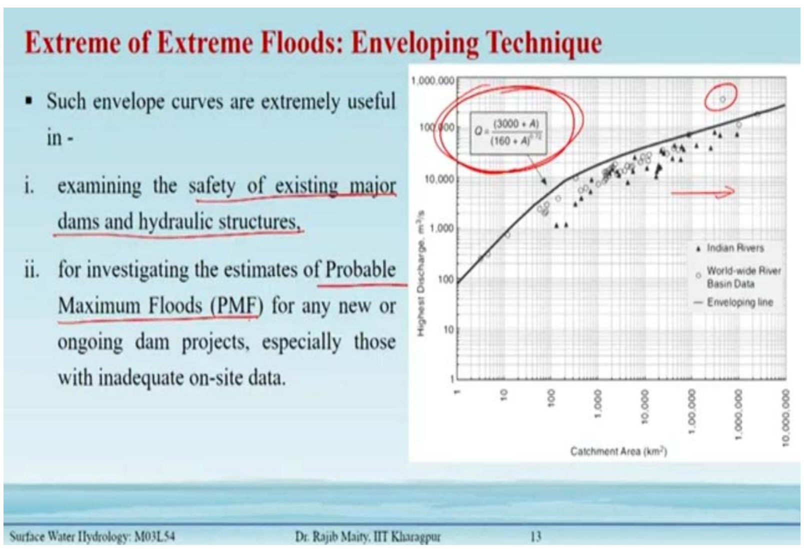


Figure 1 shows a typical envelope curve (data from 32 Indian basins and 40 other river basins spread across the globe) **(SOURCE: Subramanya, 2013)**

(Refer Slide Time: 14:21)





The enveloping line, encompassing all the plotted data points except one outlier, could be empirically expressed as:

$$Q = \frac{(3000 + A)}{(160 + A)^{0.72}}$$

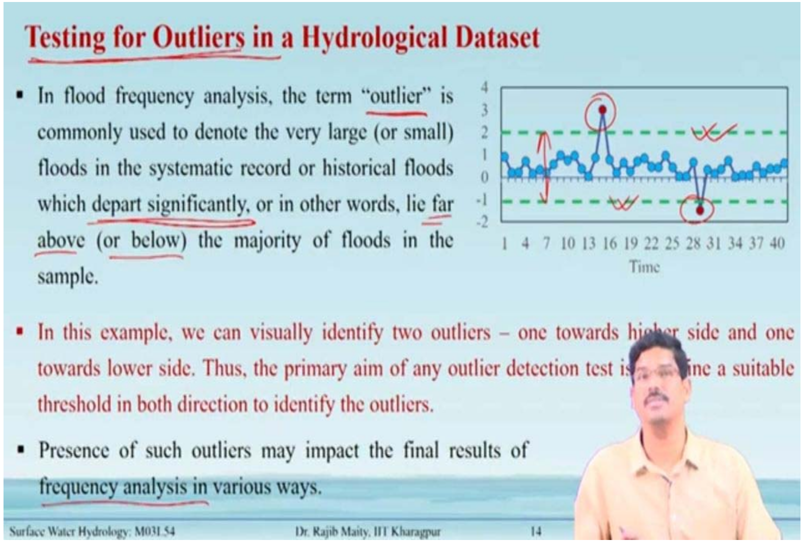
Where, Q = Highest flood discharge in m^3/s and A = Catchment area in km^2

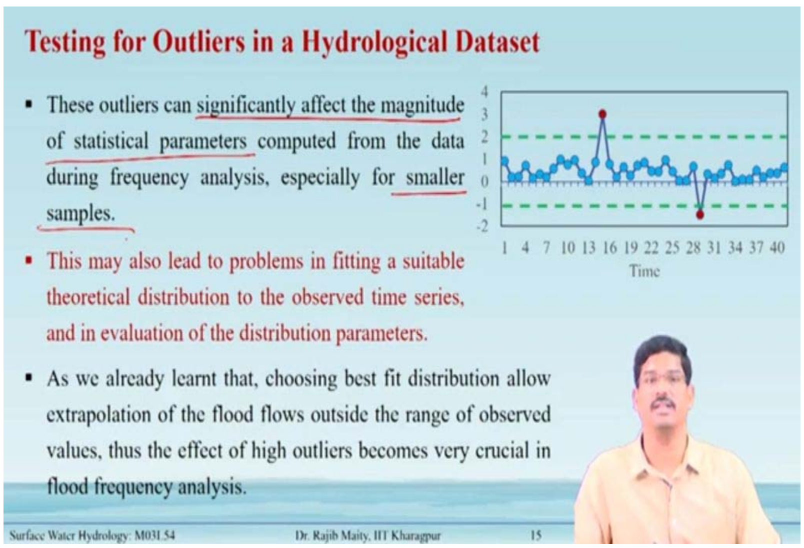
Such envelope curves are extremely useful in -

i. examining the safety of existing major dams and hydraulic structures,

ii. For investigating the estimates of Probable Maximum Floods (PMF) for any new or ongoing dam projects, especially those with inadequate on-site data.

(Refer Slide Time: 16:17)





Testing for Outliers in a Hydrological Dataset

In flood frequency analysis, the term "outlier" is commonly used to denote the very large (or small) floods in the systematic record or historical floods which depart significantly, or in other words, lie far above (or below) the majority of floods in the sample.

In fig.1, we can visually identify two outliers – one towards the higher side and one towards the lower side. Thus, the primary aim of any outlier detection test is to define a suitable threshold in both directions to identify the outliers. The presence of such outliers may impact the final results of frequency analysis in various ways.

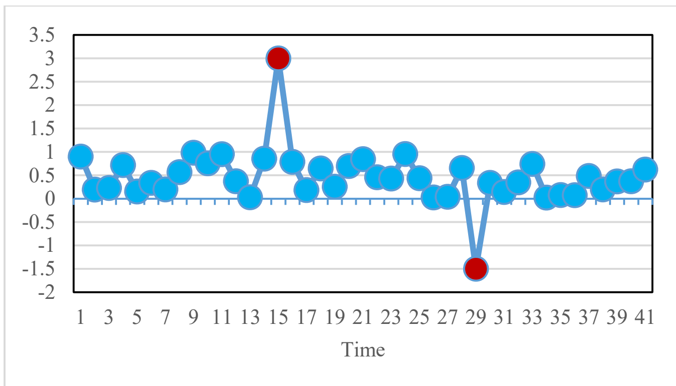
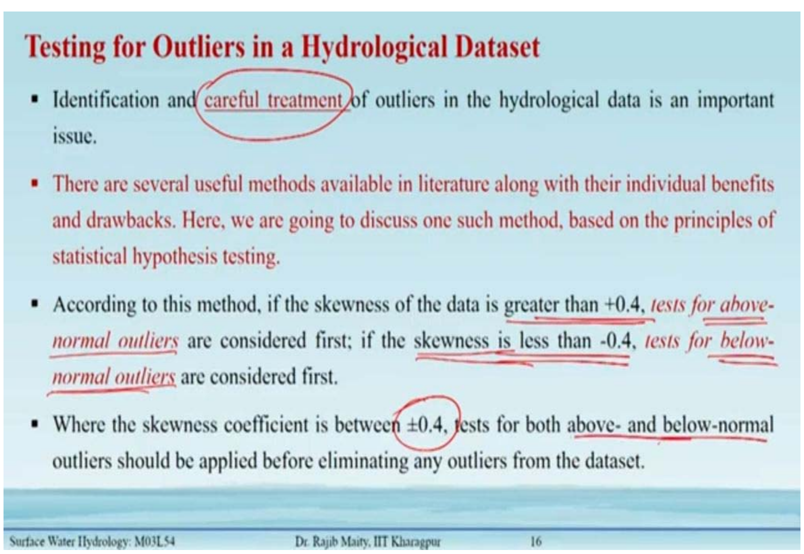


Figure 2 shows the outlier detection of a given time series

(Refer Slide Time: 18:57)



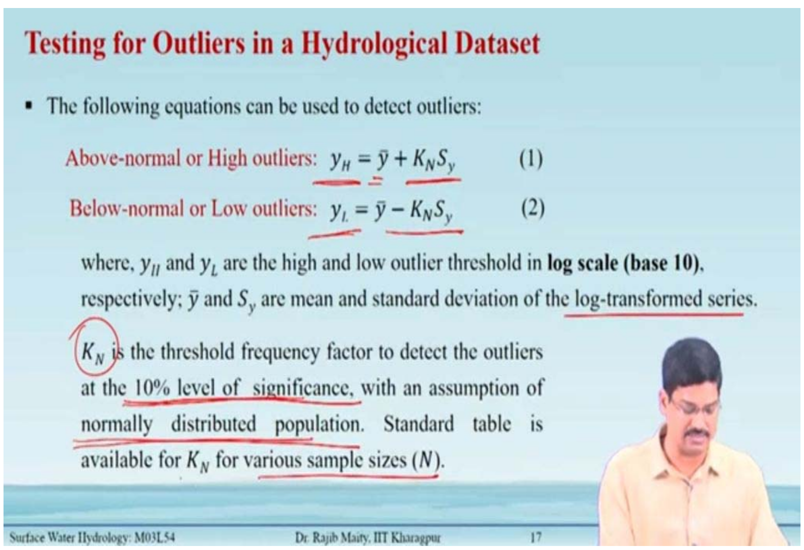
Identification and careful treatment of outliers in the hydrological data is an important issue.

There are several useful methods available in the literature along with their benefits and drawbacks. Here, we are going to discuss one such method, based on the principles of statistical hypothesis testing.

According to this method, if the Skewness of the data is greater than +0.4, tests for above-normal outliers are considered first; if the Skewness is less than -0.4, tests for below-normal outliers are considered first.

Where the Skewness coefficient is between ±0.4, tests for both above- and below-normal outliers should be applied before eliminating any outliers from the dataset.

(Refer Slide Time: 20:38)



The following equations can be used to detect outliers:

Above-normal or High outliers: $y_H = \bar{y} + K_N S_y$ (1)

Below-normal or Low outliers: $y_L = \bar{y} - K_N S_y$ (2)

Where, y_H and y_L are the high and low outlier threshold in the log scale (base 10), respectively; $\bar{y}$ and S_y are the mean and standard deviation of the log-transformed series.

K_N is the threshold frequency factor to detect the outliers at the 10% level of significance, with an assumption of a normally distributed population? A standard table is available for K_N for various sample sizes (N).

(Refer Slide Time: 21:47)

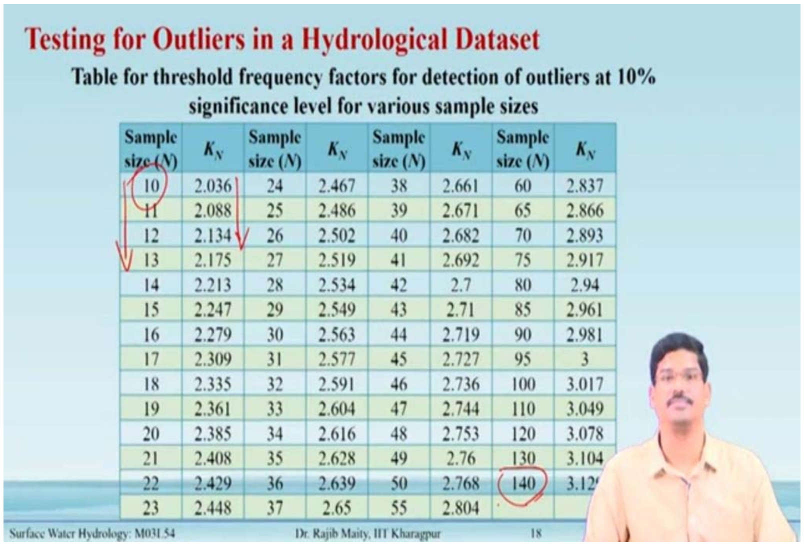
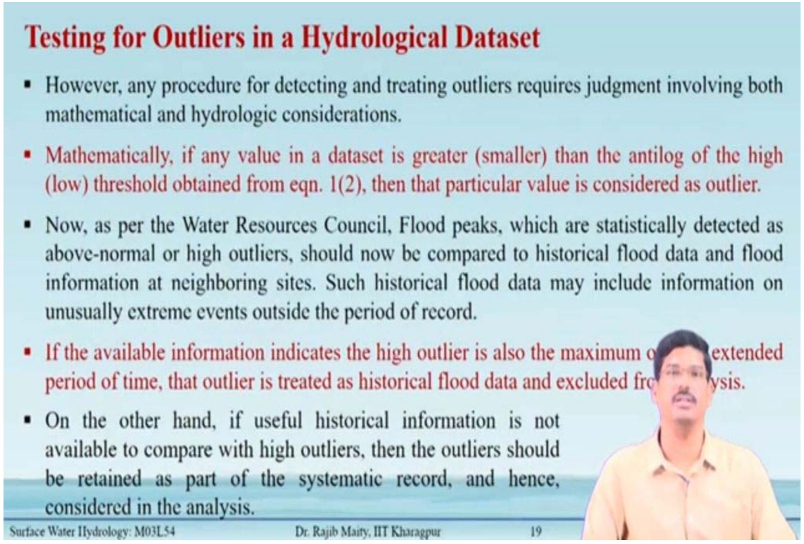


Table for threshold frequency factors for detection of outliers at a 10% significance level for various sample sizes

Sample size (N)	K_N	Sample size (N)	K_N	Sample size (N)	K_N	Sample size (N)	K_N
10	2.036	24	2.467	38	2.661	60	2.837
11	2.088	25	2.486	39	2.671	65	2.866
12	2.134	26	2.502	40	2.682	70	2.893
13	2.175	27	2.519	41	2.692	75	2.917
14	2.213	28	2.534	42	2.7	80	2.94
15	2.247	29	2.549	43	2.71	85	2.961
16	2.279	30	2.563	44	2.719	90	2.981
17	2.309	31	2.577	45	2.727	95	3
18	2.335	32	2.591	46	2.736	100	3.017
19	2.361	33	2.604	47	2.744	110	3.049
20	2.385	34	2.616	48	2.753	120	3.078
21	2.408	35	2.628	49	2.76	130	3.104
22	2.429	36	2.639	50	2.768	140	3.129
23	2.448	37	2.65	55	2.804		

(Refer Slide Time: 22:06)

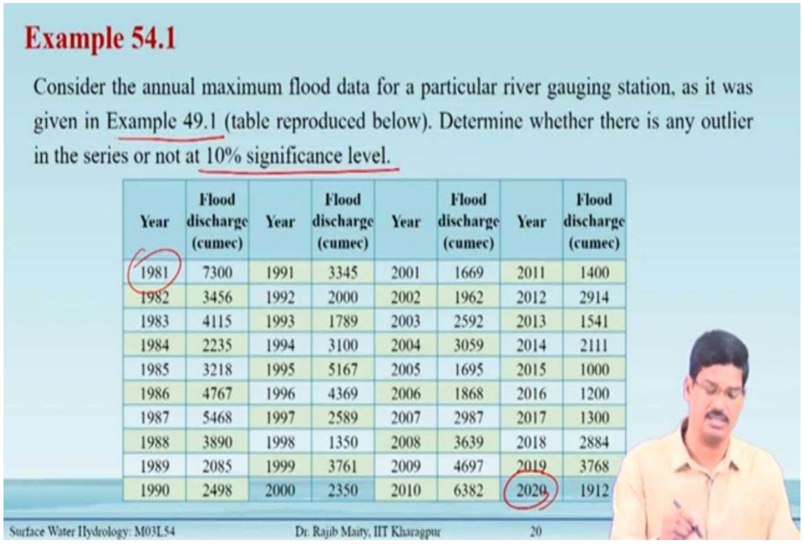


However, any procedure for detecting and treating outliers requires judgment involving both mathematical and hydrologic considerations. Mathematically, if any value in a dataset is greater (smaller) than the antilog of the high (low) threshold obtained from Eqn. 1(2), then that particular value is considered an outlier.

Now, as per the Water Resources Council, Flood peaks, which are statistically detected as above-normal or high outliers, should now be compared to historical flood data and flood information at neighboring sites. Such historical flood data may include information on unusually extreme events outside the period of record.

If the available information indicates the high outlier is also the maximum over an extended period of time, that outlier is treated as historical flood data and excluded from the analysis.

(Refer Slide Time: 24:06)

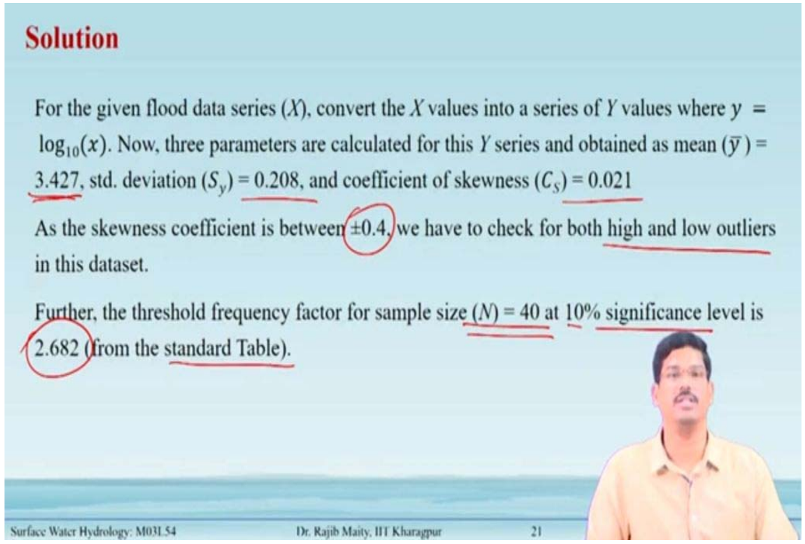


Example 54.1

Consider the annual maximum flood data for a particular river gauging station, as it was given in Example 49.1 (table reproduced below). Determine whether there is an outlier in the series or not at the 10% significance level.

Year	Flood discharge (cumec)	Year	Flood discharge (cumec)	Year	Flood discharge (cumec)	Year	Flood discharge (cumec)
1981	7300	1991	3345	2001	1669	2011	1400
1982	3456	1992	2000	2002	1962	2012	2914
1983	4115	1993	1789	2003	2592	2013	1541
1984	2235	1994	3100	2004	3059	2014	2111
1985	3218	1995	5167	2005	1695	2015	1000
1986	4767	1996	4369	2006	1868	2016	1200
1987	5468	1997	2589	2007	2987	2017	1300
1988	3890	1998	1350	2008	3639	2018	2884
1989	2085	1999	3761	2009	4697	2019	3768
1990	2498	2000	2350	2010	6382	2020	1912

(Refer Slide Time: 24:37)



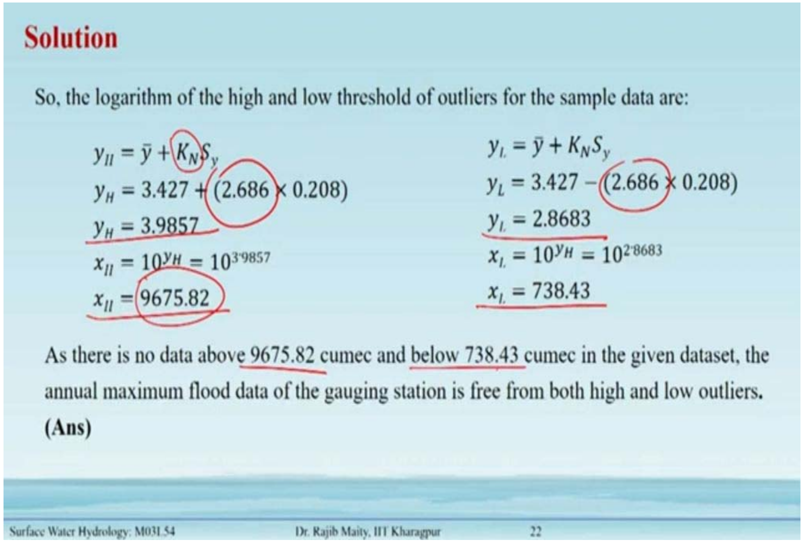
Solution

For the given flood data series (X), convert the X values into a series of Y values where$y = log_{10}(x)$. Now, three parameters are calculated for this Y series and obtained as mean ($\bar{y}$) = 3.427, std. deviation (S_y) = 0.208, and coefficient of skewness (C_S) = 0.021

As the Skewness coefficient is between ±0.4, we have to check for both high and low outliers in this dataset.

Further, the threshold frequency factor for sample size (N) = 40 at a 10% significance level is 2.682 (from the standard Table).

(Refer Slide Time: 25:21)



So, the logarithm of the high and low threshold of outliers for the sample data is:

$$y_H = \bar{y} + K_N S_y$$

$$y_H = 3.427 + (2.686 \times 0.208)$$

$$y_H = 3.9857$$

$$x_H = 10^{y_H} = 10^{3.9857}$$

$$x_H = 9675.82$$

$$y_L = \bar{y} + K_N S_y$$

$$y_L = 3.427 - (2.686 \times 0.208)$$

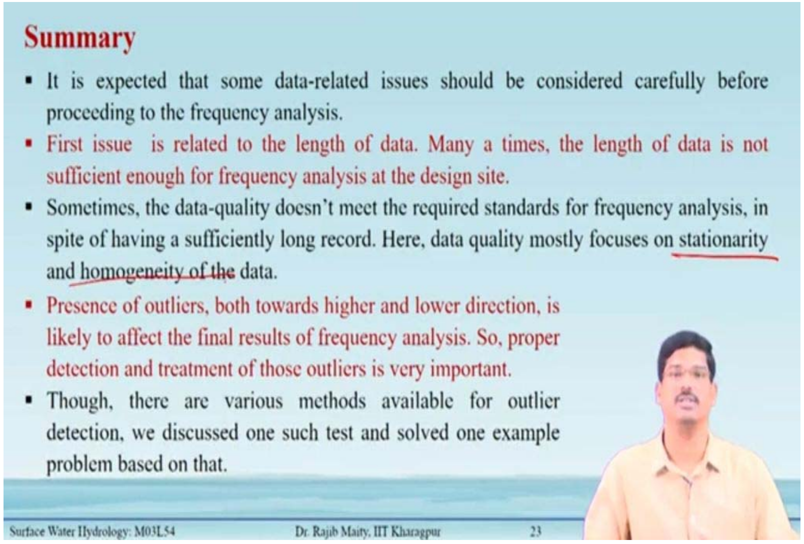
$$y_L = 2.8683$$

$$x_L = 10^{y_H} = 10^{2.8683}$$

$$x_L = 738.43$$

As there is no data above 9675.82 cumec and below 738.43 cumec in the given dataset, the annual maximum flood data of the gauging station is free from both high and low outliers. (Ans)

(Refer Slide Time: 26:05)



Summary

In summary, we learned the following points from this lecture:

- It is expected that some data-related issues should be considered carefully before proceeding to the frequency analysis.
- The first issue is related to the length of data. Many times, the length of data is not sufficient enough for frequency analysis at the design site.
- Sometimes, the data quality doesn't meet the required standards for frequency analysis, despite having a sufficiently long record. Here, data quality mostly focuses on stationarity and homogeneity of the data.
- The presence of outliers, both in higher and lower directions, is likely to affect the final results of frequency analysis. So, proper detection and treatment of those outliers are very important.
- Though there are various methods available for outlier detection, we discussed one such test and solved one example problem based on that.

Reference

- Subramanya, K. (2013). Engineering Hydrology, 4th edition. Tata McGraw-Hill Education, New Delhi, India