

**Course Name: Watershed Hydrology**

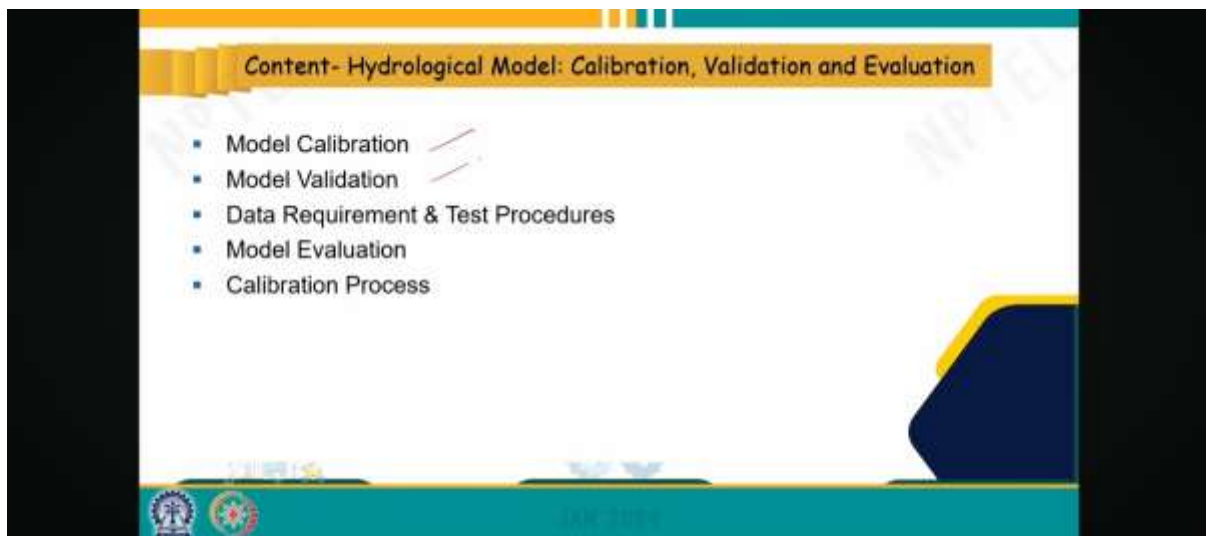
**Professor Name: Prof. Rajendra Singh**

**Department Name: Agricultural and Food Engineering**

**Institute Name: Indian Institute of Technology Kharagpur**

**Week: 08**

**Lecture 38: Hydrological Model: Calibration, Validation and Evaluation**



Hello friends, welcome back to this online certification course on Watershed Hydrology. I am Rajendra Singh, a professor in the Department of Agriculture and Food Engineering at the Indian Institute of Technology Kharagpur. We are in module 3, this is lecture number 3, and the topic is Hydrological Model Calibration, Validation and Evaluation. So, in this lecture, we will talk about model calibration, model validation, data requirements, and test procedures, as well as model evaluation and the calibration process. Starting with calibration, model calibration is the process of estimating model parameters. Here, model predictions or output are compared with the observed data for a given set of showed conditions or for a set of fixed model parameters. So, as we also discussed earlier, in the modelling protocol there are two major steps: calibration and validation.

## MODEL CALIBRATION AND PARAMETER

- **Calibration** is the process of estimating model parameters
- Model predictions (**output**) are compared with the **observed data**
  - ✦ For a given set of assumed conditions, i.e., for a **fixed set of model parameters**

```

graph TD
    MP[Model Parameters] --> PP[Physical Parameters]
    MP --> PrP[Process Parameters]
    
```

- Represent the **physical properties of the catchment**
- Usually **measurable**
- Example: catchment area, surface slope
- Represent **catchment characteristics**
- Usually **non-measurable**
- Example: the average depth of water storage capacity, Manning roughness coefficient
- There are some physical parameters, such as the hydraulic conductivity and porosity, which are measurable in theory but difficult to measure in practice, and hence are often calibrated

Now, we are talking about calibration. Here, we change the model parameters, tune them, and then run the simulation. Once the simulation is completed, we compare the model prediction, the simulated output, with the observed output. If we are satisfied, fine. Of course, we have to have some performance criteria, which we will discuss later, but if we are not satisfied, then obviously, we go back and tune the model parameters and then run the simulation again. This process continues unless and until the model is satisfied with the performance of its model.

So, this process is a very vital step because that is where you really tune and correct the model parameters so that you are able to simulate the processes correctly. And as far as model parameters go, there are two types: physical parameters and process parameters. Now, physical parameters represent the physical properties of the catchment, and these are usually measurable. That means you can measure these parameters; for example, catchment area surface slope. So, obviously, with a little bit of experimental effort, you can measure these parameters. On the other hand, process parameters represent the catchment characteristics, and these may not be measurable. Usually, they are non-measurable.

For example, the average depth of water storage capacity, Manning's roughness coefficient—I mean you can do laboratory experiments and get these values, but basically, you cannot measure in the field conditions what the Manning's roughness coefficient is at a particular place. It is not really possible to go and measure. Similarly, there are some physical parameters such as hydraulic conductivity and porosity which are measurable in theory, but difficult to measure in practice, and hence are often calibrated. That is simply because, as we can see, hydraulic conductivity and porosity are soil characteristics. And in our earlier discussion, we have seen that the soil characteristics change every few meters or few hundred meters within the catchment or a basin or a watershed. So, obviously, in order to get a correct value of hydraulic conductivity or porosity for the entire basin, you have to have hundreds of experiments conducted, which is very time-consuming and, of course, resource-intensive.

## MODEL CALIBRATION AND PARAMETER

- **Calibration** is the process of estimating model parameters
- Model predictions (**output**) are compared with the **observed data**
  - ❖ For a given set of assumed conditions, i.e., for a **fixed set of model parameters**

**Model Parameters**

↓

**Physical Parameters**

- Represent the **physical properties of the catchment**
- Usually **measurable**

Example: catchment area, surface slope

**Process Parameters**

- Represent **catchment characteristics**
- Usually **non-measurable**

Example: the average depth of water storage capacity, Manning roughness coefficient

- There are some physical parameters, such as the hydraulic conductivity and porosity, which are measurable in theory but difficult to measure in practice, and hence are often calibrated

So, that is why instead of measuring them, we take the knowing of the soil texture. We know the range of hydraulic conductivity or porosity, and then we try to calibrate these parameters so that we get the right value of the parameter. That is the process of in the ah, I mean, I mean, in the process of this calibration ah, we change the parameters as we discussed, and then we measure, we compare the model results with the observed field data, and then we say whether our model is calibrated or not.

So, that is the calibration process. Then comes model validation, which is the next step and of course, a crucial step in assessing the accuracy and reliability of the hydrological model. So, in the calibration process, we have calibrated the model by playing with the model parameters and then obviously, we also want to see that when we are not changing the model parameter, how the model performs and that step is validation. So, it involves running a model using input parameters determined during the calibration process. Once we say the model is calibrated, that means we say that our parameter values are calibrated, and then we do not change those calibrated model parameters and then we run the model with a different set of data ah during validation.

## MODEL VALIDATION

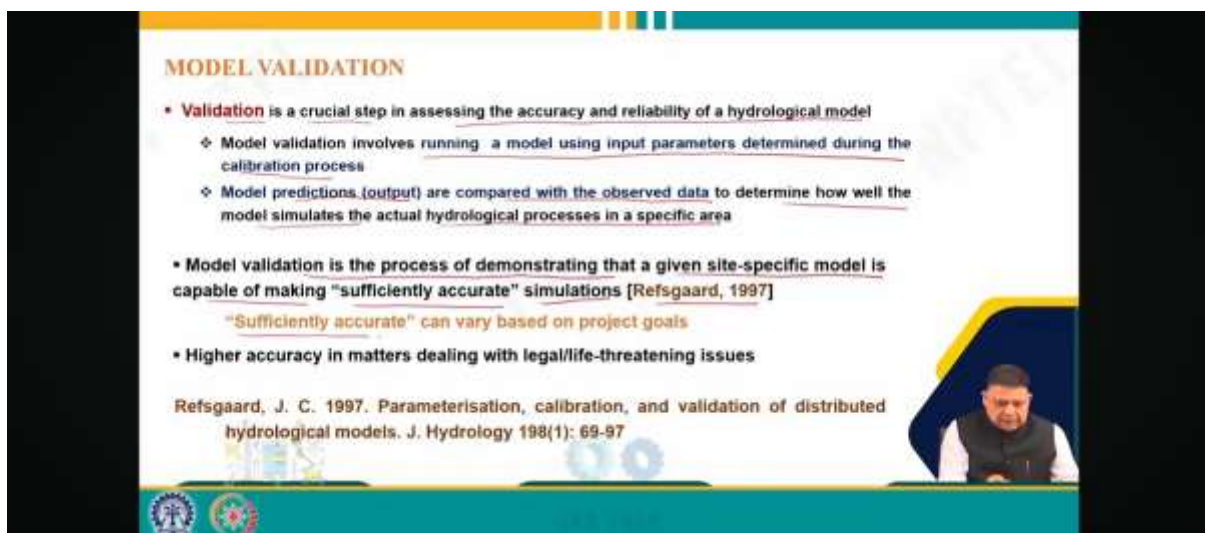
- **Validation** is a crucial step in assessing the accuracy and reliability of a hydrological model
  - ❖ Model validation involves running a model using input parameters determined during the calibration process
  - ❖ Model predictions (**output**) are compared with the **observed data** to determine how well the model simulates the actual hydrological processes in a specific area
- Model validation is the process of demonstrating that a given site-specific model is capable of making "**sufficiently accurate**" simulations [Refsgaard, 1997]
  - "**Sufficiently accurate**" can vary based on project goals
- Higher accuracy in matters dealing with legal/life-threatening issues

Refsgaard, J. C. 1997. Parameterisation, calibration, and validation of distributed hydrological models. *J. Hydrology* 198(1): 69-97

And the model prediction or output is compared with the observed data to determine how well the model simulates the actual hydrological process in a specific area. So, of course, this process can you run the model then you compare the data the simulated data with the observed

data and really see whether your model is doing good or bad. And as I said that we are not able to not allowed to change the model parameter due to validation run. So, if you are not satisfied then of course, you have to go back to calibration change the model parameters again get a new set of parameters or take the more lender data of ah lend the data set and then you try to recalibrate your model and come back and this process continues unless you say you are able to say that my model is calibrated as well as validated. Ah So, model validation is a process of demonstrating that a given site-specific model is capable of making sufficiently accurate simulations and that comes from Refsgaard 1997.

And when we say sufficiently accurate it varies ah from ah based on project goals. So, of course, you need higher accuracy in matters dealing with legal alive-threatening issues. For example, if you are trying to model dam break ah where ah if you are wrong in your estimation then if by chance something happens to them then obviously, the downstream ah area life will be threatened of course, the materials will be threatened and might bring legal issues. So, obviously, you have to be very careful in such cases, but otherwise also you would always like your model to perform as good as possible. And ah this is the reference of the paper Refsgaard J.C. 1997 parameterization calibration validation of distributed hydrological models. It has taken been taken from general of hydrology this is a very ah important parameter to understand the parameterization calibration validation process of hydrological model.



**MODEL VALIDATION**

- **Validation** is a crucial step in assessing the accuracy and reliability of a hydrological model
  - ❖ Model validation involves running a model using input parameters determined during the calibration process
  - ❖ Model predictions (output) are compared with the observed data to determine how well the model simulates the actual hydrological processes in a specific area
- Model validation is the process of demonstrating that a given site-specific model is capable of making "sufficiently accurate" simulations [Refsgaard, 1997]
  - "Sufficiently accurate" can vary based on project goals
- Higher accuracy in matters dealing with legal/life-threatening issues

Refsgaard, J. C. 1997. Parameterisation, calibration, and validation of distributed hydrological models. *J. Hydrology* 198(1): 69-97

Now, taking calibration validation together and trying to understand once again, emphasizing the difference. So, in both cases, the model is run with input parameters, and model predictions are compared with the observed data; that process remains the same. So, what is the difference between calibration and validation? The difference is here, that if we talk about the change of model parameters, then in calibration, you can do that.

Yes, you are allowed to change the model parameters during calibration, but if we talk about validation, you are not allowed to change the model parameters; that is the major difference. During the calibration process, you change the model parameters and keep on changing until you are satisfied, until you get a better good match between model simulated and observed data. And once you say that my model is calibrated, that means you are satisfied with the model parameters; then you are not allowed to change them. Then you go for a validation run and there you are not allowed to change the model parameters; that is the significant difference. Then, of course, as the input data set is concerned, the data set must be different; that means

you cannot use the same data set on which you have calibrated the model for validation; you have to have a different set of data.

**CALIBRATION & VALIDATION**

- In both cases, model is run with input parameters and model predictions (output) are compared with the observed data
- Difference between calibration and validation

Functions	Calibration	Validation
Change of model parameters	Yes	No
Input data set	MUST be different	

Same data set (fully or partially) should not be used for both calibration & validation

The same data set, fully or partially, should not be used for both calibration and validation; that is a big no-no. So, obviously, you have to get a fresh set of data when validating your model. So, this is the thin difference: during the calibration process, you are allowed to change the model parameters; in validation, you are not and it's advisory that the data set used during calibration should not be used during validation; you have to have a fresh data set.

**DATA REQUIREMENT & TEST PROCEDURES**

- Data Requirement

Observed data requirement	Calibration	Validation
Yes	Yes	Yes

Typically, longer data set is needed for calibration than validation

- Recommended test procedures

Split-Sample Test

Proxy-Catchment Test

Differential Split-Sample Test

So, that is the major difference between these two. As far as data requirement goes, obviously, we need observed data for both calibration as well as validation. A typically longer dataset is needed for calibration than for validation. And of course, there are different procedures. If we come to recommend test processes, there are three different procedures which are recommended. The simplest one is known as split sample test, then we have a differential split sample test, and lastly, we have proxy catchment test. We will discuss each of these tests one by one, and we will also talk about the length of data required or the type of data required.

**DATA REQUIREMENT & TEST PROCEDURES**

- Test procedures
  - Split-Sample Test
    - one period of observations is used in model calibration and one or more separate periods are used in model validation
    - Example

Data period	Calibration	Validation
1991-2010	1991-2005	2006-2010
	1991-2002	2003-2006
		2007-2010

Now, coming to the split sample test, as the name itself suggests, the total sample data is split into two parts: one period of observation used for model calibration and one or more separate periods are used in model validation.

For example, let us say we have 20 years of data from 1991 to 2010. Then we may use, say, 1991 to 2005, that is 15 years of data for calibration, and then run validation on 2006 to 2010. So, that is one possibility. The other possibility is that we use 1991 to 2002, that is 12 years of data for calibration, and then we can use 2003 to 2006 and 2007 to 2010, two different sets of data for two different validations. So, that is also a possibility. So, typically, 80-20 is generally recommended—80 percent for calibration and 20 percent for validation—but there is no hard and fast rule. Depending upon the length of data, you can decide what length is appropriate for calibration and validation, and of course, you can also have multiple validations. So, typically, this is a typical picture: if you have 1991 to 2010 data, then you might use the first year of data, 1991, 1992, as a warm-up period. This warm-up period is basically to give a better initial condition to the model because, as you can understand, if your model starts in January, say, in the conditions in January when there is not much moisture, or the same model starts in July when there is a lot of rainfall occurring.

So, of course, the conditions will be very different and then the starting point being different, simulation will also be affected. So, that is why typically what we do is that we use one year of data for a warm-up period. It is also in modelling what is also known as hot start in some literature. You will see the term hot start instead of warm-up period. So, after, and of course, this data is not used for the results are not used in analysis. And then after the warm-up period, then you run 1993 to 2002 data. You can run for calibration and then you can have two independent validations: 2003 to 2006 and 2007 to 2010. And then, of course, you compare in calibration also in validation also you compare the model simulated results with the field day observed results and then you decide how your model is doing. So, that is a split sample test. The next test is a differential split sample test where data is, in the previous case, we saw the data was simply divided based on the year.

**DATA REQUIREMENT & TEST PROCEDURES**

• **Differential Split-Sample Test**

1  
Data is divided according to rainfall rate (or some other variable) in an attempt to show that the model has some general validity

2  
For example, Monsoon and Non-monsoon periods for rainfall, or High flows & Low flows for streamflow

Model is then cross-validated.  
For example  
Calibrate for high flows and validate for low flows and vice-versa

So, you have 20 years of data; you use 15 years for calibration, 5 years for validation without going into the characteristics of the data, but in a differential split state, data is divided according to some particular characteristics of the variable, say rainfall rate or some other variable in an attempt to show that the model has some general validity. So, you know. It is a little bit complicated test compared to split sample test. So, for example, we may say that we will run the simulation for monsoon and non-monsoon periods, two different, will simulate for monsoon, calibrate for monsoon then validate for non-monsoon or we can say that we will calibrate for high flows and validate for low flows in the case of streamflow. So, depending on what kind of variable you are using, a specific characteristic of the variable is chosen and based on that, data is divided into two different sets.

So, monsoon versus non-monsoon or high flow versus low flow values are typically then modelled and cross-validated. What is done is that we calibrate for high flows and validate for low flows, and vice versa. That means you calibrate for low flows and then validate for high flows. So, you are checking both ways that you are calibrating for high flows set of parameters then validating for low flows, and you are calibrating for low flows and then validating for high flows. In both cases, you are comparing the observed and simulated results and then establishing the general validity of your model. So, this is a differential split sample test. The last one is even more challenging, that is called the proxy catchment test, and here we use data for two different catchments to show that the model has even greater general validity.

Of course, we take two data sets from two different catchments, and it involves calibrating the model against data for one catchment and running a validation test for the other catchment. Obviously, these two catchments we choose should be in a hydrologically homogeneous region, their hydrological characteristics should be the same. They could be neighbouring catchments, but obviously, we calibrate for one and validate for the other catchment data. We can also do cross-validation where we calibrate for the second catchment and then validate for the first catchment and then compare the results. So, that is also a possibility.

### DATA REQUIREMENT & TEST PROCEDURES

• Proxy Catchment Test

- 1 Uses data for two catchments to show that the model has even greater general validity
- 2 Involves calibrating the model against data for one catchment and then running a validation test for the other catchment
- 3 The proxy catchment is treated as ungauged at the model estimation stage, but with some observations of discharge and perhaps other variables which are available for the evaluation of the model predictions

The proxy catchment is treated as ungauged at the model estimation stage, but with some observations of discharge and perhaps other variables which are available for the evaluation of model prediction. Now, this proxy catchment test is really helpful because sometimes we may have to use our model for ungauged catchments.

So, even if you treat the proxy catchment as ungauged, although you have the major data, but still, you treat it ungauged and calibrate and validate it. Then, it gives you confidence that if you have a catchment which is ungauged where you do not have sufficiently long data, then also you calibrate your model for hydrologically homogeneous catchment and then you apply the same model into an ungauged catchment in order to stimulate the desired variable. So, that is even a greater general validity test of the model. These are three different tests having three different characteristics and of course, the split sample test is most commonly used, but other tests can also be used. Now, once we have calibrated and validated that means, then of course, during the entire process we compare the observed simulation with the simulated value. So, obviously, we have to evaluate the model and so, model evaluation techniques come into the picture.

### MODEL EVALUATION

```

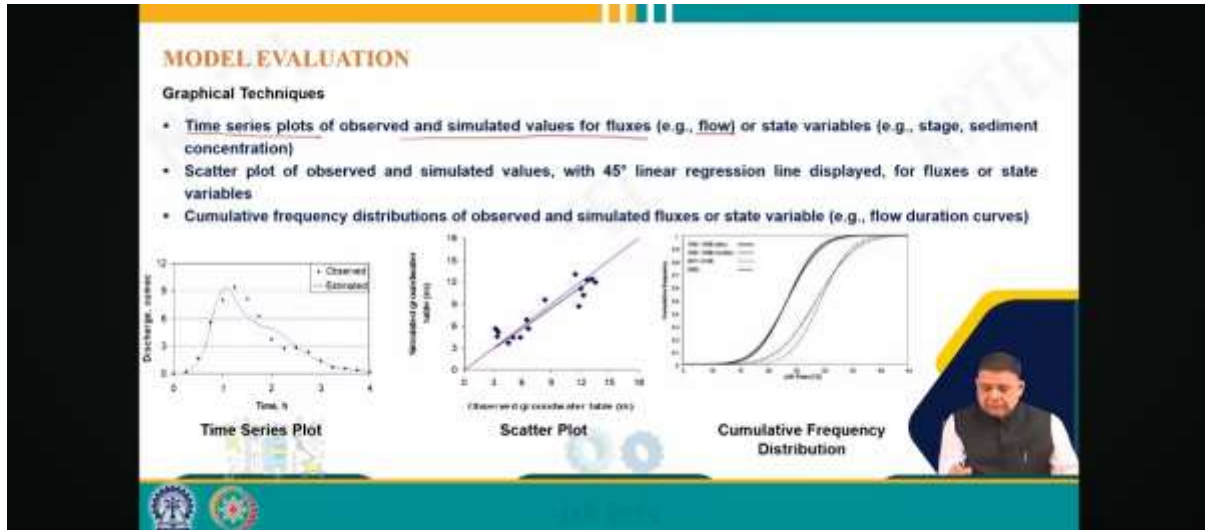
graph TD
    A[Model Evaluation Techniques] --> B[Graphical Techniques]
    A --> C[Quantitative statistics]
    B --> D[Time Series plots]
    B --> E[Scatter plots]
    B --> F[Cumulative frequency distributions]
  
```

- Graphical plots provide a visual comparison of simulated and measured constituent data
- Gives the first overview of model performance

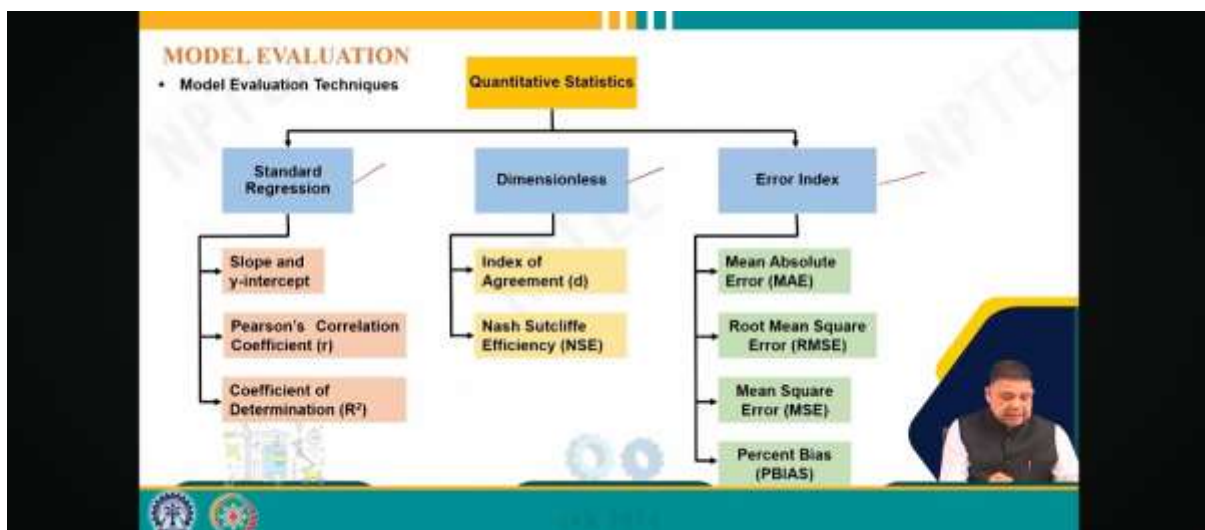
Model evaluation techniques can be broadly classified into graphical techniques and quantitative statistics. If we talk about graphical techniques, then there are three possibilities: we go for time series plots, we can plot scatter plots or we can also have cumulative frequency



distribution curves. Graphical plots provide a visual comparison of simulated and measured constituent data. So, of course, if you plot the observed versus simulation data in the same graph, it gives you a clear picture of whether your model is able to simulate the desired pattern of the variable or not, whether it is able to match the peaks, low simulations, and medium. Suppose flow we are talking then whether it is able to meet that or not, all these things you can graphically visualize. So, of course, graphical technique gives us the first overview of model performance.



So, looking at the plot, we can really feel whether our model is doing good or bad. So, that gives you the first impression about the model. So, when talking about graphical techniques, it is said that we plot time series plots of observed and simulated values of fluxes. For example, flow or state variables like sediment concentration or whatever. Here, it is the time series plot where discharge observed and estimated discharge are plotted on the same curve. Observed is represented by dots and the thin pink line represents the estimated value.



As you can see here, in the rising limb of the hydrograph, the model is doing pretty well. Of course, the peak is not well simulated; there is some time gap and the simulated peak is coming earlier than the observed one. Also, the initial part of the falling limb of the hydrograph is not doing pretty well in this portion. So, at the first glance, you get an idea about how your model is performing. Then the next type of plot is a scatter plot where we plot the simulated and

observed variables. Here, groundwater table is plotted, simulated groundwater table versus observed groundwater table. Of course, this is plotted against the 45-degree line, the one-to-one line. Ideally, if your model is performing perfectly, then all data points will be on the 45-degree line.

**MODEL EVALUATION**

- Quantitative Statistics
  - Standard Regression

**Slope and y-intercept**

- Slope indicates the relative relationship between simulated and measured values
- y-intercept indicates the presence of a lag or lead, or that the data sets are not perfectly aligned
- Slope of 1 and y-intercept of 0 is ideal, showing perfect match

$y = mx + c$

m is the slope of the line

y-intercept (c)

So, obviously, that is not feasible in a physical model. So, all your data should be as close as possible to the 45-degree line. If you are scattered well on both sides of the 45-degree line, then we say, "Okay, my model is doing well." Otherwise, if there are scatters, then you may say, "Okay, my model is not doing well," and you will go back and change the model parameters. The third thing is, of course, the cumulative frequency distribution. Again, here you compare the observed and simulated cumulative frequency distribution of the variable. So, whatever variable, it could be flow, it could be any flux also, and then you can compare and really see whether your model is doing good or bad at first glance. And of course, after graphical techniques, after evaluating through graphical techniques, and once you are satisfied graphically, then comes the quantitative statistics, of course and when we talk about quantitative statistics, then again, we can have three different types of quantitative statistics.

**MODEL EVALUATION**

- Quantitative Statistics
  - Standard Regression

**Pearson's Correlation Coefficient (r)**

- Describe the degree of collinearity between simulated and measured data
- r is an index of the degree of linear relationship; r = 0 shows no linear relationship where ± 1 shows perfect linear relationship
- r is oversensitive to high extreme values (outliers)

Simulated data

Measured data

$r = 0.9$

The standard regression, dimensionless and error index. And within each group, there are a number of indices. For example, in standard regression, we have slope and y-intercept, Pearson's correlation coefficient, and coefficient of determination. And the important or

popular dimensionless ones are the index of agreement and natural efficiency which is the most popular quantitative statistic and under error index, we have mean absolute error, root mean square error, mean square error, percent bias. So, there are many, out of which percent bias and root mean square error are pretty common. And of course, you know that coefficient of determination is pretty commonly used.

**MODEL EVALUATION**

- Quantitative Statistics
  - Standard Regression

**Coefficient of Determination ( $R^2$ )**

**Coefficient of Determination ( $R^2$ )**

- $R^2$  describes the proportion of the variance in measured data explained by the model
- $R^2$  ranges from 0 to 1, with higher values indicating less error variance, and typically values greater than 0.5 are considered acceptable
- $R^2$  is oversensitive to high extreme values (outliers)

Simulated data

Measured data

$R^2 = 0.9$

So, of course, there are some very popular ones, but these are the various possibilities as far as quantitative statistics go. So, starting with regression, we have standard regression that includes slope and y-intercept. So, here it is like fitting a best-fit curve or best-fit line and saying it is a straight line if you fit then of which equals to  $mx$  plus  $c$ , where  $m$  provides the slope of the line and  $y$  is the intercept. And slope indicates the relative relationship between simulated and measured value, that is the angle, and y-intercept indicates the presence of a lag or lead or that the data sets are not perfectly aligned. So, obviously, for ideal conditions, the slope will be 1, that is a 1-to-1 line, and the y-intercept will be 0, which will show the perfect match.

**MODEL EVALUATION**

- Quantitative Statistics
  - Dimensionless

**Index of Agreement (d)**

**Index of Agreement (d)**

- It provides a relative model evaluation
- Varies between 0 and 1, with 1 being the perfect value
- Sensitive to extreme flows due to squared difference
- Relatively high values (more than 0.55) of  $d$  may be obtained even for poor model fits, leaving only a narrow range for model calibration

$$d = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (|P_i - \bar{O}| + |O_i - \bar{O}|)^2}$$

$O$  and  $P$  are observed & predicted data;  
 $\bar{O}$  is the observed average;  
 $n$  is the number of observations

So, obviously, your deviation from slope 1 and intercept 0 will tell you how inaccurate you are or how your model is doing. So, that is one indication of qualitative statistics. Of course, we commonly use Pearson's correlation coefficient which describes the degree of collinearity

between simulated and observed data. So, obviously, we plot simulated versus observed data and then we try to get the best-fit line and then calculate the value of R. And R is an index of the degree of linear relationship. R equal to 0 shows no linear relationship, whereas  $\pm 1$  shows a perfect linear relationship, that is for a linear curve but it can also be a power or exponential relationship and then, R value, of course, plus minus 1, is always ideal, and R is oversensitive to high or extreme values. So, if you have outliers, very high values, then of course, they will impact the value of R significantly. Then we have the coefficient of determination, R square, which describes the proportion of the variance in the major data explained by the model. It ranges between 0 and 1, with a higher value indicating less error variance and typically values greater than 0.5 are considered acceptable. And of course, the ideal value is 1; R square value is 1, and R square is also oversensitive to high extreme values, which is quite obvious. So, obviously, when we say R square of 0.5, that means the value of R is almost 0.7. So, that means, a very good explanation of the simulated observed match between simulated and observed value.

**MODEL EVALUATION**

Quantitative Statistics

- Dimensionless

Nash Sutcliffe Efficiency (NSE)

- Varies between  $-\infty$  and 1, with 1 being the perfect value
- Value lower than zero indicates that the observed mean value would have been a better predictor than the model
- Sensitive to high flows due to squared difference
- Overestimation of the model performance during peak flows and an underestimation during low flow conditions
- Very Popular – recommended by ASCE (1993)

$$NSE = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2}$$

O and P are observed & predicted data;  
 $\bar{O}$  is the observed average;  
 n is the number of observations

ASCE. 1993. Criteria for evaluation of watershed models. *J. Irrigation Drainage Eng.* 119(3): 429-442

The coefficient of determination is also a pretty popular one. Coming to dimensionless types, as we said, there are two major ones. The first one is the index of agreement D which provides a relative model validation and is calculated using this relationship.

$$D = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (|P_i - \bar{O}| + |O_i - \bar{O}|)^2}$$

Where:

- $O_i$  represents the observed values.
- $P_i$  represents the predicted values.
- $\bar{O}$  denotes the mean of the observed values.
- $n$  is the number of observations.

This metric provides a relative measure of model performance, with values closer to 1 indicating better agreement between observed and predicted data.

So, from here, 1 minus this will give us the value of D or the index of agreement and it is varied between 0 and 1, with 1 being the perfect value. And of course, this is sensitive to extreme

flows due to the square difference. So, obviously, you can see that because if there is an error in the extreme values and because the square term is being used. So, obviously, the value of D will be affected.

**MODEL EVALUATION**

Quantitative Statistics

- Error Index → Quantify the deviation in the units of the data of interest

Mean Absolute Error (MAE) | Mean Square Error (MSE)

Root Mean Square Error (RMSE)

- Useful because these express error in "Units", with 0 being the ideal value for all three
- RMSE and MAE are usually recommended
- RMSE and MAE values less than half the standard deviation of the measured data are considered low

Relatively high values, more than 0.65 of D, may be obtained even for poor model fits, leaving only a narrow range for model calibration. That is the one drawback of the index of agreement that even a high value of 0.65 does not indicate a good model fit. So, because of the formulation, you get a very high value.

**MODEL EVALUATION**

Quantitative Statistics

- Error Index

Percent Bias (PBIAS) | Deviation of Flow Volume (Dv)

**PBIAS**

- Measures the average tendency of the simulated data to be larger or smaller than their observed counterparts
- Ideal value is 0
- Positive values indicate model underestimation bias and negative values indicate model overestimation bias

$$PBIAS = \frac{\sum_{i=1}^n (O_i - P_i) * 100}{\sum_{i=1}^n O_i}$$

O and P are observed & predicted data; n is the number of observations

**Dv**

- Calculated in a fashion similar to PBIAS - also recommended by ASCE (1993)

ASCE, 1993. Criteria for evaluation of watershed models. *J. Irrigation Drainage Eng.* 119(3): 429-442

So, unless you get a value greater than 0.7 or more, then only can you say that ok my model is satisfactory. So, one has to be really careful about the value of D and then claiming whether the model is doing good or bad. And of course, the most commonly used dimensionless model evaluation statistic is Nash-Sutley efficiency, which is calculated using this relationship:

$$1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{n}$$

Where:

- $O_i$  represents the observed data at each index  $i$ ,
- $P_i$  represents the predicted data at each index  $i$ ,

- $O$  is the total observed data,
- $\bar{O}$  is the mean of the observed data,
- $n$  is the number of observations.

So, obviously, observed and predicted values you are using for comparison or for calculating NSC value. And its value varies from minus infinity to 1, with 1 being the perfect value.

So, if you get an NSC equal to 0.1 or close to 0.1, that is, close to 1. Sorry, if you get an NSC value of 1, close to 1, that is 0.9 or so, then obviously, your model is doing pretty good. Values lower than 0, that means negative values, indicate that the observed mean value would have been a better predictor than the model. So, if you get an NSC value at minus, then that shows that your model is good for nothing. Instead of using the model, if you had used the average observed value, probably that would have been a better predictor. So, obviously, you have to get a positive value of NSC and it is close to 1, possible.

**MODEL EVALUATION**

- Adoption of Model Evaluation Technique(s)

Recommendation

At least one from each group → Dimensionless statistic, Error Index statistic, one graphical technique

- Time-Series Plot,  $R^2$ , Nash-Sutcliffe Efficiency, RMSE and Percent Bias are the popular choices
- Additional information such as the standard deviation of measured data may also be included

And of course, this is also sensitive to high flows due to square difference, which you can see that if observed predicted values are at high values are different, then obviously, a large square is being used so, obviously, NSC value will be affected. And overestimation of the model performance during peak flow and underestimation during low flow, that is a typical characteristic of NSC. And it is very popular, as I have already mentioned, and it is recommended by the American Society of Civil Engineers in 1993. Actually, the American Society of Civil Engineers in 1993 formulated a committee to recommend the criteria for evaluation of watershed models because even today, if you find a number of papers, people use different statistics. So, it was happening that time also in the early '90s also when the modelling studies came into work.

### CALIBRATION PROCESS

```

graph TD
    Calibration[Calibration] --> Manual[Manual]
    Calibration --> Automatic[Automatic]
    Manual --> ManualText[Model parameters are adjusted by the modeller  
Trial and Error Process - time consuming  
Difficult to determine the "best fit" or to determine a clear point indicating the end of the calibration process  
Different modellers may obtain different results.]
    Automatic --> AutomaticText[The purpose is to find those values of the model parameters that optimise (minimise or maximise) the numerical value of the objective function]
    
```

**Equifinality**

- Multiple parameter sets providing equally good or acceptable model outputs
- One of the main sources of uncertainty in hydrological modelling

So, that is where the ASC formulated committee. So, to so, this committee recommended 3 and they said any 2 of the 3 should be used and NSC is one of them. NSC is one of the dimensionless statistics that was recommended by this committee for evaluation of watershed models and that is why you find that almost every single paper will report NSC value in today's research. Then we come to error index and under error index already we listed that is we have mean absolute error, root mean square error, mean square error. And of course, this quantifies the deviation in the units of data of interest that is an important thing that in this case units of data are used. That means, if it is stream flow then obviously, you will be RMC suppose you are using root mean square error then of course, you will write that the RMC of flow is so much cubic meter per second.

So, in this case unit comes into picture. So, it gives you an idea about and you know the data the how much your flow range is. So, it gives you a numerical idea of how your model is doing good or bad. So, that is the significance of this useful because these express error in units with 0 being the ideal value of all 3. So, whether you use mean absolute error RMC or MSC if your RMC value is 0 for flow 0 cusec that means, you are absolutely perfect in simulating the observed data, but of course, it is hypothetical. And RMC and mean absolute error are usually recommended and RMC and ME value less than half the standard deviation of the major data are considered low.

### CALIBRATION PROCESS

```

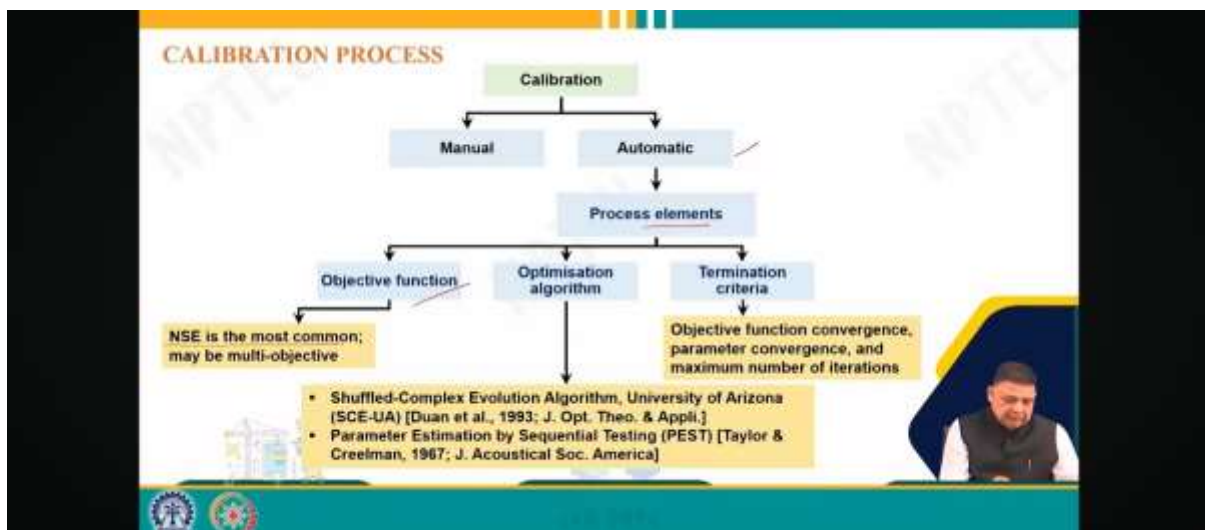
graph TD
    Calibration[Calibration] --> Manual[Manual]
    Calibration --> Automatic[Automatic]
    Automatic --> ProcessElements[Process elements]
    ProcessElements --> ObjectiveFunction[Objective function]
    ProcessElements --> OptimisationAlgorithm[Optimisation algorithm]
    ProcessElements --> TerminationCriteria[Termination criteria]
    ObjectiveFunction --> ObjectiveText[NSE is the most common; may be multi-objective]
    OptimisationAlgorithm --> AlgorithmList[Shuffled-Complex Evolution Algorithm, University of Arizona (SCE-UA) [Duan et al., 1993; J. Opt. Theo. & Appl.]  
Parameter Estimation by Sequential Testing (PEST) [Taylor & Creelman, 1967; J. Acoustical Soc. America]]
    TerminationCriteria --> TerminationText[Objective function convergence, parameter convergence, and maximum number of iterations]
    
```

So, obviously, when you take the data, we have already discussed the data analysis, you take the data of course, you calculate the mean and standard deviation first. So, obviously, if the RMS value, which you are calculating from the model output by comparing the model output and the field observed data. So, if the RMS value is less than half of the standard deviation then that means, it is an acceptable value. So, this error index has the advantage of using the units. Now, coming to the error index, the other two are the P bias which measures the average tendency of the simulated data to be larger or smaller than their observed counterparts.

And this is how it's used: simply percent deviation, basically percent bias, the percent deviation we use  $O_i$  minus  $P_i$  by  $O_i$ , and it is expressed in percent, the ideal value of 0. So, obviously, the bias should be 0 and then that will be ideal, but it's known that it is not possible for a physical model. And positive values indicate model underestimation bias, and negative values indicate model overestimation bias because if you get a positive value that means, the observed data are higher than the predicted one so that means, the model is underestimating. And if you get a negative value then it is just vice versa, then your predicted values are more than the observed value, that means, your model is overpredicting. And similar to P bias, there's another one which is deviation in flow volume, which is calculated in a fashion similar to P bias which is also recommended by ASC.

So, I've already mentioned two out of the three recommendations by the ASC. They recommended a deviation of flow volume, NSC and two others. This DV, calculated as P bias is more commonly used. However, the calculations, procedures, and meanings remain the same, making it a widely used method. Now, we've observed numerous model evolution values, both graphical and quantitative. We have dimensionless values error indices and graphical techniques. Essentially, we recommend utilizing at least one of these three: dimensionless, error index, and graphical techniques.

Consequently, time series plots can display natural wave efficiency graphically, dimensionless values, and error indices like RMC and percent bias. These are common statistical choices. Additional information such as the standard deviation of major data may also be included to reflect the model's performance. During calibration and validation, it's expected to provide a graphical plot, like a time series plot, and use an error index, such as percent bias or RMC along with dimensionless statistics like natural efficiency. These must be adopted and reported when presenting results.





Moving on to the calibration process, there are two main methods: manual calibration and automatic calibration.

So, both of these calibrations are possible, and in the manual calibration model, parameters are adjusted by the modeler, and it is a trial-and-error process. It is time-consuming and difficult to determine the best fit or to determine a clear point indicating the end of the calibration process. Basically, because if you change any one of the model parameters, then obviously, your results will change, and you may be doing slightly better than the previous one. So, you will be tempted to go for energy change and so on. It is a very time-consuming process because the modeler is changing parameters one by one. So, if a model has a large number of parameters, of course, it will take a long time to calibrate the model manually.

Another problem is that different modelers may obtain different results. So, if n number of models are trying to calibrate the same model with the same data set, everyone might come up with a different combination, and this is a problem referred to as equifinality. That is multiple parameter sets providing equally good or acceptable model outputs, which is a problem with manual calibration. It is one of the main sources of uncertainty in hydrological modelling because different combinations of parameters can give you a good result at another set of different combinations. So, that means there is always uncertainty in your model results; there is no conformity, there is no best, something like it is the best parameter set. I mean that is the problem. On the other hand, we have automatic calibration where the purpose is to find those values of the model parameters that optimize, that is minimize or maximize the numerical value of a particular objective function which we use.

So, automatic calibration involves three different process elements. First, you must have an objective function. We may use NSC (Normalized Scalar Coefficient), and naturally, we aim to maximize NSC. Alternatively, we can opt for a multi-objective function.

Next, you need two optimization algorithms. Two popular choices are the Shuffled Complex Evolution Algorithm (SCEUA), developed at the University of Arizona, hence its name, and Parameter Estimation by Sequential Testing (PEST), the publication reference for which is available. These algorithms, SCEUA and PEST, are commonly used.

Additionally, termination criteria are necessary. This could involve setting a maximum number of iterations, such as 500 or 1000 runs, and accepting the resultant values. Alternatively, convergence of the objective function can serve as a termination criterion. For instance, if the NSC value decreases by less than 1% between iterations, the process can be terminated. The same principle applies to parameter convergence—if there is no change in parameter values after a 1% adjustment, the process can be deemed satisfactory.

So, of course, while automatically calibrating, one has to have these three process elements: an objective function an objective optimization algorithm you have to use and of course, you have to set termination criteria. So, with this, we come to the end of this lecture. We have talked about the calibration validation and evaluation and also the calibration process like manual or automatic calibration. Thank you very much. Please give your feedback and also raise your doubts or questions which we will be happy to answer.



Thank you very much.