Course Name: Watershed Hydrology

Professor Name: Prof. Rajendra Singh

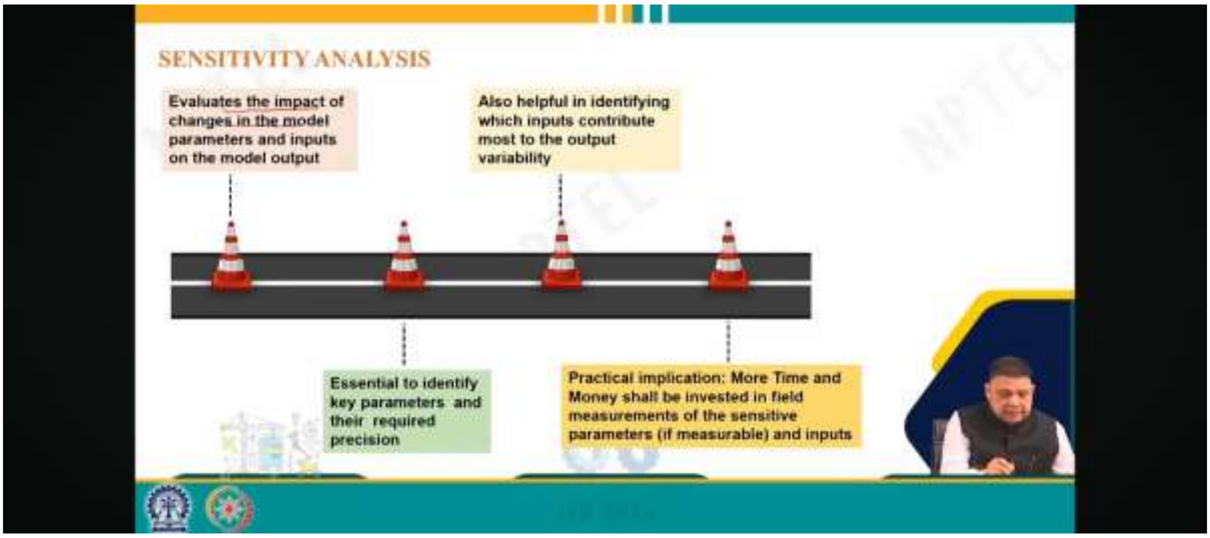
Department Name: Agricultural and Food Engineering

Institute Name: Indian Institute of Technology Kharagpur

Week: 08

Lecture 39: Sensitivity Analysis and Machine Learning in Hydrology

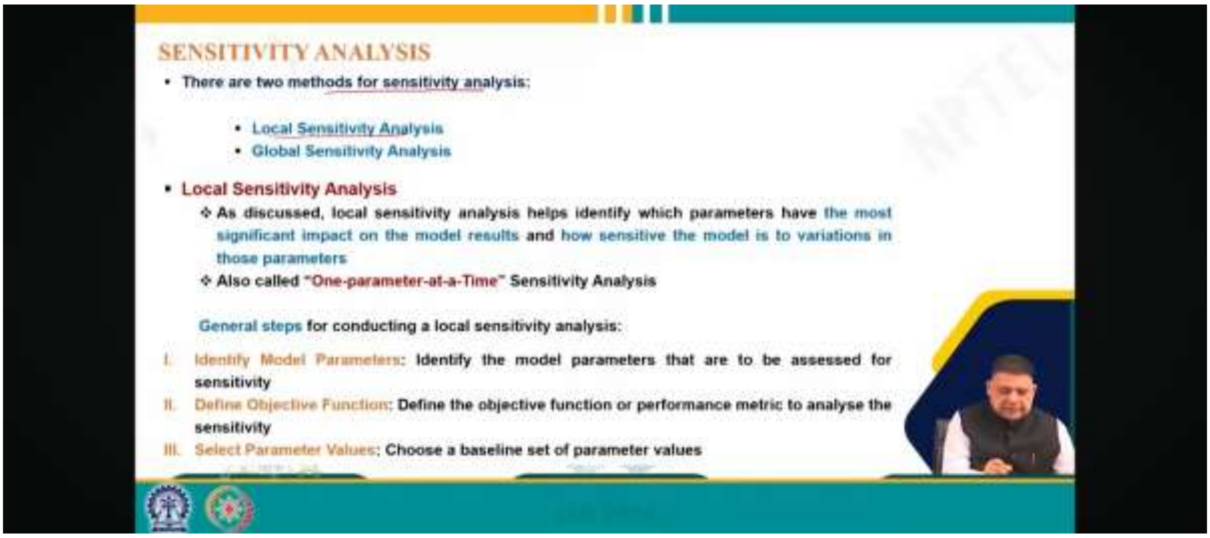
Hello friends, welcome back to this online certification course on Watershed Hydrodynamic Hydrology. I am Rajendra Singh, a professor in the Department of Agriculture and Food Engineering at the Indian Institute of Technology Kharagpur. We are in module 8, this is lecture number 4 and the topic is Sensitive Analysis and Machine Learning in Hydrology. So, we will have two parts. In the first part, we will talk about sensitivity analysis, and then we will shift into machine learning in hydrology. So, this lecture will cover sensitivity analysis. We will introduce sensitivity analysis, solve problems on sensitivity analysis, then move into machine learning. We will go through deep learning and classification of machine learning.



Now, talking about sensitivity analysis, basically, it evaluates the impact of changes in the model parameters and input on the model output. That means, if your model parameters or your input parameters change, then how your model output is affected is what we analysed under sensitivity analysis. And it is essential to identify key parameters and their required precision. So, obviously in this process, we try to identify key parameters to which our model is very sensitive, meaning changes in those parameter values will significantly affect our model results. Therefore, we have to be more precise about the values of those parameters.

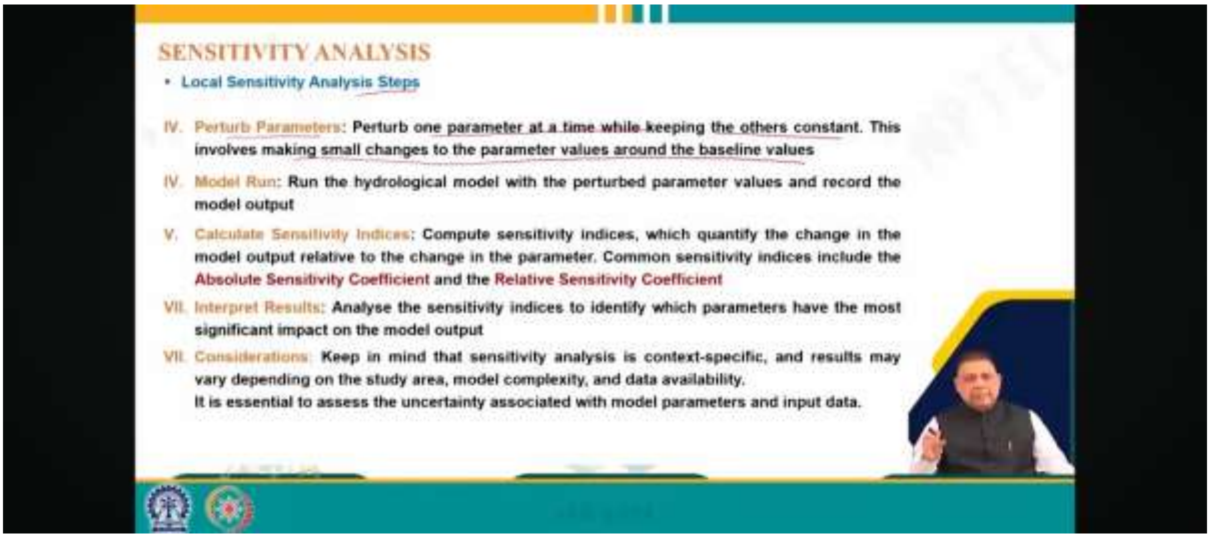
Similarly, it helps in identifying which inputs contribute most to the output variables. So, obviously, among various inputs, we can also identify which input is more significant. The practical implication of sensitivity analysis is that more time and money shall be invested in field measurements of sensitive parameters if measurable and inputs. That simply means that because we only have limited resources and suppose a model has 10 parameters and then it has

4 or 5 inputs. And so, if we go through sensitivity analysis, we try to find out okay these 2 parameters are more sensitive and also these 2 inputs.



So, whatever resources we have, we should try to collect precise information on those parameters and those inputs because you know that their values will be making an impact on the model output. So, obviously the entire exercise of sensitivity analysis, the practical implication is that we can identify important parameters and inputs from the modelling perspective. And then we should focus on investing more resources in getting more precise values of those parameters and variables. There are 2 methods of carrying out sensitivity analysis: the local sensitivity analysis and the global sensitivity analysis. In the local sensitivity analysis, local sensitivity analysis helps identify which parameters have the most significant impact on the modulators and how sensitive the model is to variations in those parameters.

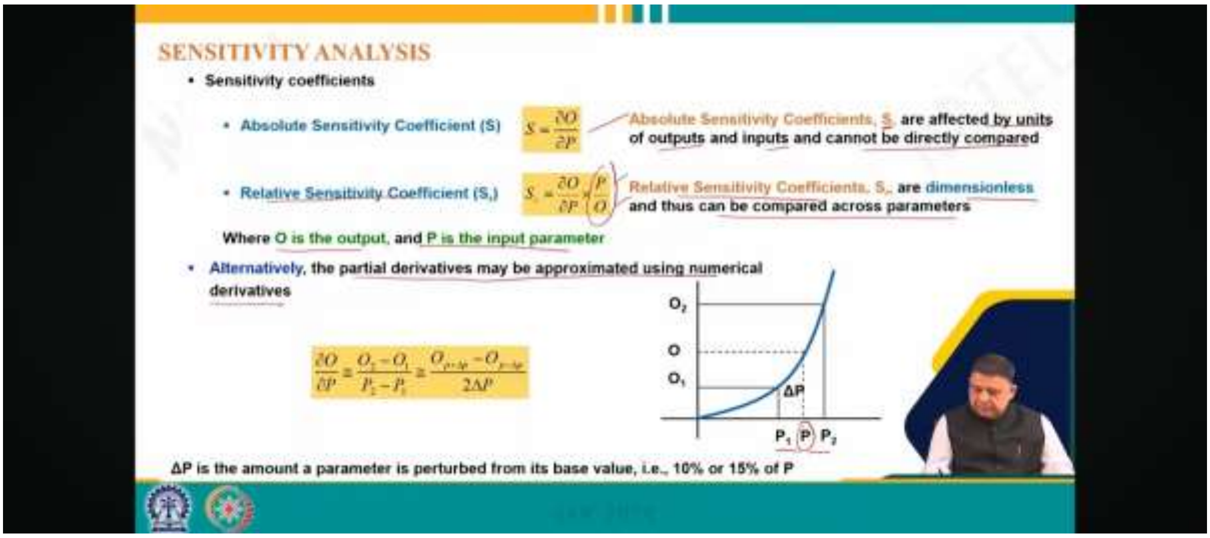
So, the same thing we discussed in sensitivity analysis is done here. It is also called one parameter at a time sensitivity analysis. So, in this case if you have 10 parameters in a model, then the modeler will change one parameter at a time while keeping the other 9 parameters constant and then analysed the model. The sensitivity of the output to changes in that particular parameter will be analysed, and the same is done for all parameters. Generally, steps for conducting a local sensitivity analysis involve identifying the model parameters, meaning what the model parameters are, and we may choose, if there are 10, we may select 5 that seem to be more sensitive.



So, we can carry out the model sensitivity analysis on those model parameters. Then, of course, we have to define the objective function, meaning we have to have some criteria or performance matrix based on which we will analyse the sensitivity. And, of course, then we will select the parameter value, choose a baseline set of parameter values. So, obviously, you have to have a base value of the parameter and the range over which the parameter varies so that you can change the values and then do the sensitivity analysis or carry out the sensitivity analysis. Local sensitivity analysis steps include perturbing the parameters, which means the next steps are to perturb one parameter at a time while keeping the others constant.

This involves making a small change in the parameter values around the baseline. So, if you know the baseline of a particular parameter, then you change it by, say, 5 percent, 10 percent, or 1 percent, depending upon how much time or resources you have and how sensitive you want your analysis to be based on that you will change. And then you go for model run, the hydrological model with the perturbed parameter value, and record the model output. And then you calculate the sensitivity indices, so based on the outputs you get, you calculate the sensitivity indices. And common sensitivity indices include the absolute sensitivity coefficient and relative sensitive coefficient, and we will see the definition of these a little later.

Then, of course, you interpret results that is analysed the sensitivity indices to identify which parameters have the most significant impact on the model output. So, by comparing the sensitivity of different parameters, you will say that this parameter is more sensitive. And of course, you have to keep in mind that sensitivity analysis context-specific results may vary depending on the study area, model complexity, and data availability. So, of course, there is a sensitive analysis, not a general analysis for a particular model. Of course, you can always say that this model parameter is important, but the relative importance of different parameters could change from one basin to another, from one dataset to another. So, that is why every time it is recommended that you carry out the sensitivity analysis.



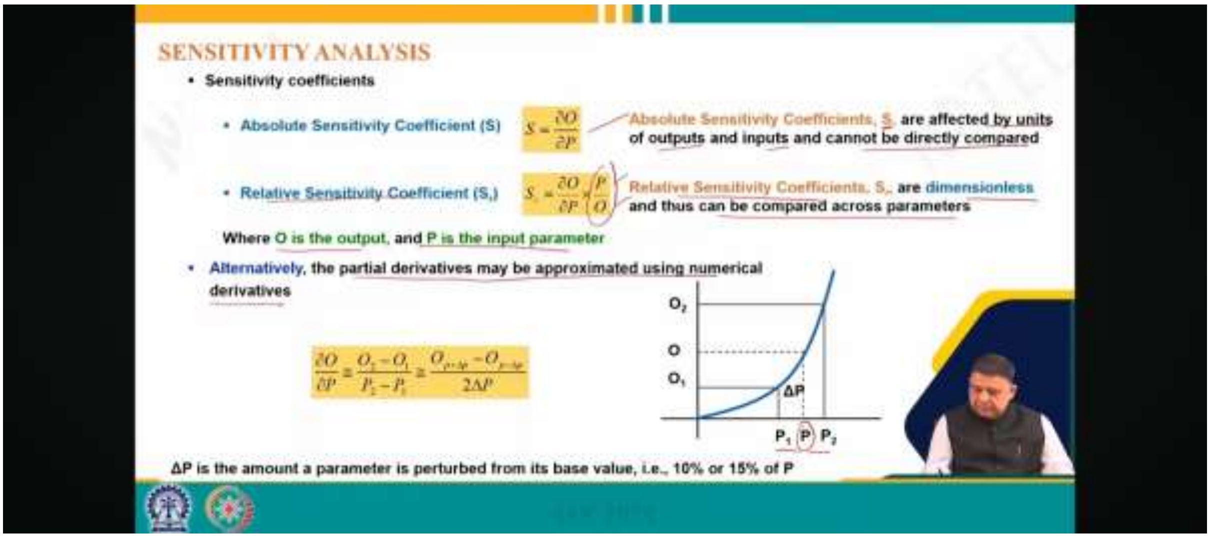
Now, coming to sensitivity, sensitive coefficients, we have absolute sensitivity coefficient, which is $(\frac{\delta O}{\delta P})$ and relative sensitivity coefficient, that is

$$\left(\frac{\delta \boldsymbol{O}}{\delta \boldsymbol{P}}\right) \times \frac{\boldsymbol{P}}{\boldsymbol{O}}.$$

where O is the output and P is the input parameter. So, absolute sensitivity coefficients, which are designated by S, are affected by units of outputs and inputs and cannot be directly compared. So, obviously, if you have discharge and you have a say, for example, the diameter of a particular pipe, and if you have many roughness's, their units and values will be significantly different. So, obviously, through absolute sensitivity coefficients, you cannot compare them relatively. On the other hand, with relative sensitivity coefficients, they are dimensionless because we multiply the value $\frac{P}{O}$. Here we are calculating $\frac{\delta O}{\delta P}$, but when we multiply by $\frac{P}{O}$, it becomes dimensionless and can be compared across parameters. Alternatively, you can use partial derivatives or alternatively the partial derivatives may be approximated using numerical derivatives.

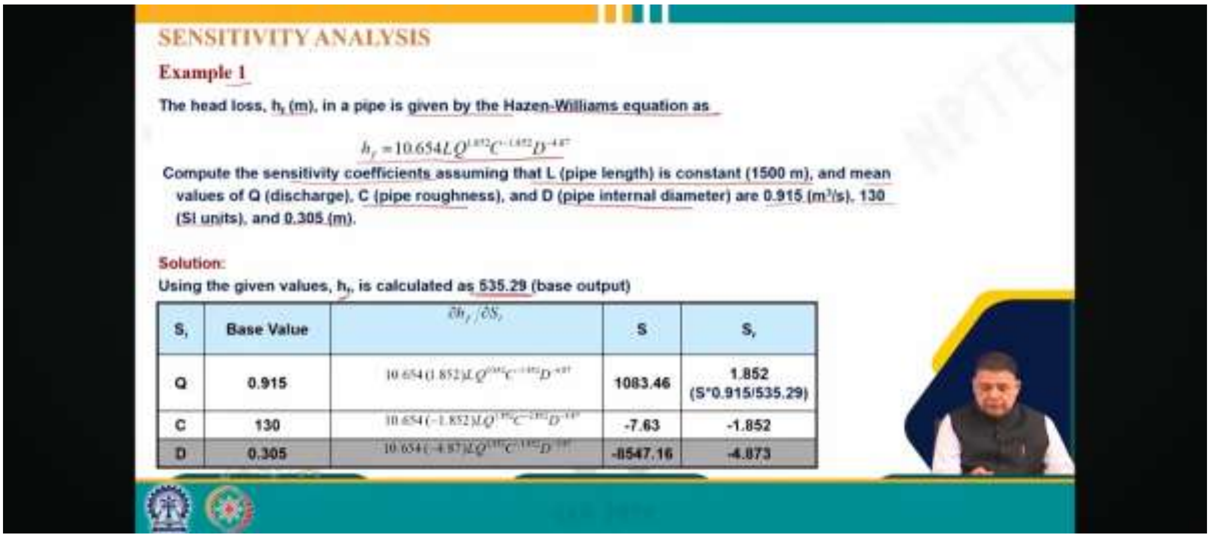
So, basically, here your parameter value is P. You can perturb that by 5 percent, 10 percent, or 15 percent on either side. For each of those, you will get the output value, say P_1 and P_2at the output.

$$\frac{\Delta \boldsymbol{O}}{\Delta \boldsymbol{P}} = \frac{\boldsymbol{O_2} - \boldsymbol{O_1}}{\boldsymbol{P_2} - \boldsymbol{P_1}}$$



So, this is how we may approximate the sensitive coefficient. Let us take an example: the head loss H_f in a pipe is given by the Hagen Williams equation, which is given here. Compute the sensitive coefficients, assuming that L is constant; its value is 1500 meters, a mean value of Q discharge C, that is pipe roughness and D, pipe internal diameter, at 0.915 cubic meters per second, 130 SI units and 0.035 meters.

So, obviously, the first thing we will do is knowing the base values of Q, C, L, and D, which are given here. We will calculate the Hf value. So, the base value or base output is 535.29, simply putting the values of 1500, 0.915, 130 and 0.305 in this equation. We get the value of H_f as 535.29 which is referred to as the base output actually. And then we have three different variables here: Q, C and D, and their base values are already given: 0.915, 130 and 0.305. And of course, then we differentiate this Hagen Williams equation, which is referring to this particular say Q.

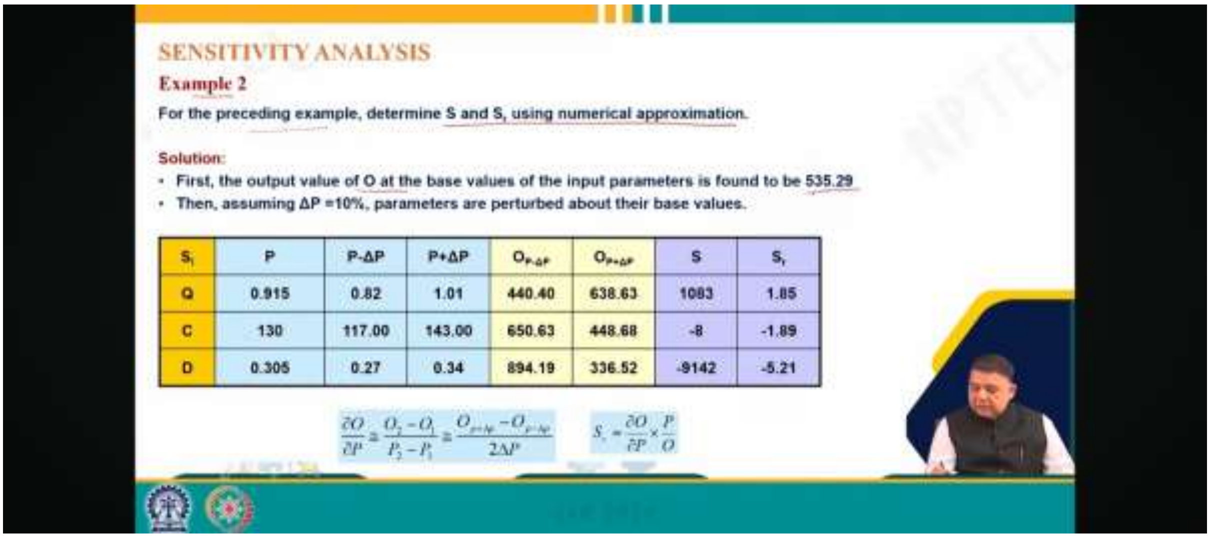


So, $\frac{dH}{dQ}$ of course, if you do this is what you will get, and then you put the value of L, Q, C and D in this and then you will get the value of S, that is 1083.56, 46. But in order to calculate the relative, we have to multiply by the base input and base output, which we already have these two values. Thus, we will get the relative coefficient. Similar analysis we will do for C, that means, we will do $\frac{dH}{dC}$ and then we will get this function. We will put the values, calculate the value of S, calculate the value of SR and the same analysis will be done for D also.

And of course, as you can see that as we said that S depends on the units so obviously, the value varies drastically and you cannot compare them, but in the case of SR, it is dimensionless.

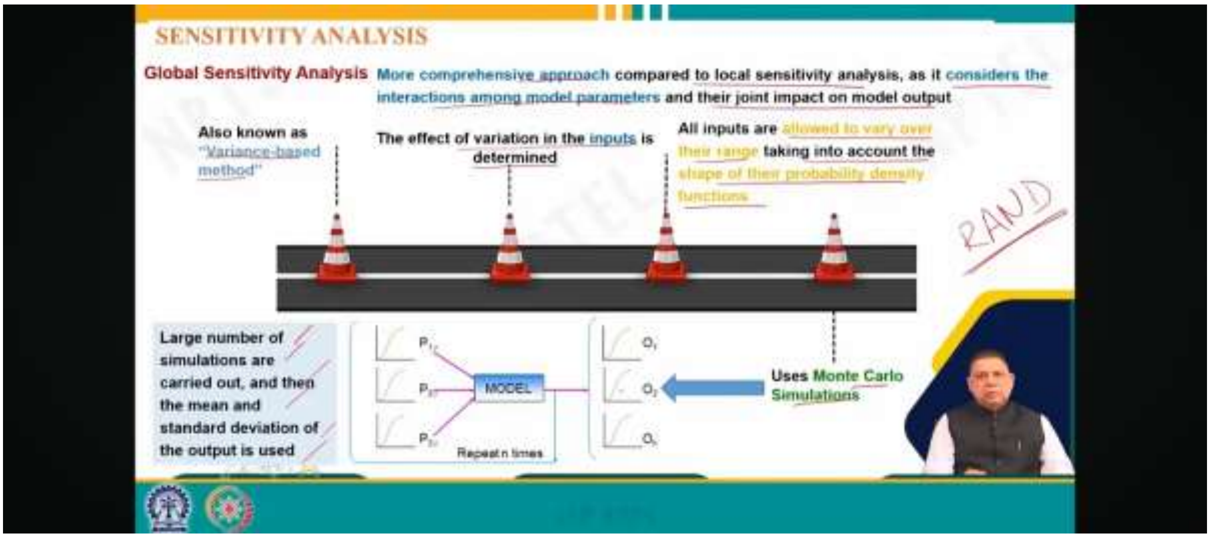
So, you can compare across the parameters, and from here we can simply see that D is the most sensitive parameter of the T. So, Q C and D so, in Hagen Williams equation and this particular problem pipe internal diameter is the most sensitive parameter. So, that is the answer of this question. You can take another example that is for the preceding example determine S and S R using numerical approximation. So, as we said that partial differentiation, we can also numerically approximate. So, first the output value of O, the base value which is calculated with the same, will remain the same. Then assuming delta P is 10 percent, parameters are perturbed about their base value.

So, Q C and D, their base values are known. So, we have 10 percent plus and 10 percent minus, that is 0.82 and 1.01, and you will calculate O P minus delta P, P plus delta P, then we calculate S and S R as usual. So, basically del O by del P is nothing but this equation. Using this, we can calculate del O by del P, that is the value of S and then S R we can calculate. This is S and this is S R. Of course, we will use the base input and base output value, and here also we get the same answer, that is D is the most sensitive parameter. The numerical value is slightly different, but we get the same answer.



Then we come to global sensitivity analysis, which is a more comprehensive approach compared to local sensitive analysis. It considers the interaction among model parameters and their joint impact on a model output. So, in this case, as we saw in the local sensitivity analysis, it is one parameter at a time. So, we only change one parameter, other parameters remain constant, but in the global sensitivity, all parameters change. So, their interaction also comes into the picture. This is also called global sensitivity, also called variance-based method, and the effect of variation of the inputs is determined, which is quite obvious, that is the sensitivity analysis, and all inputs are allowed to vary over their range, taking into account the shape of their probability density function.

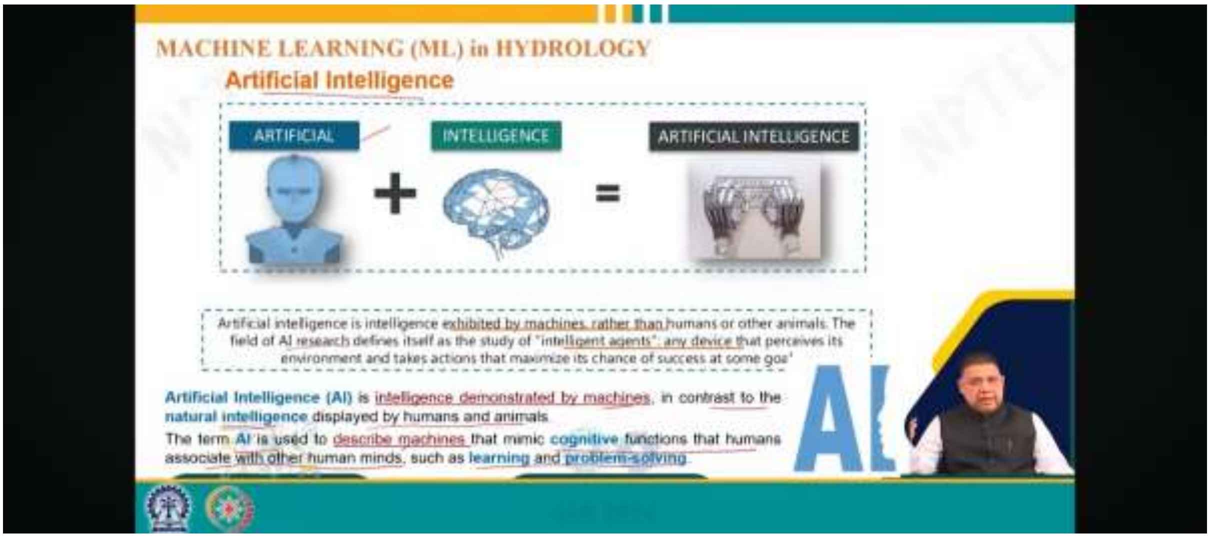
So, basically, we have already established that we know each parameter and their respective range. Within this framework, we can utilize Monte Carlo simulation or any random number generation technique. As mentioned earlier in Excel, there is a rand function for generating random numbers. You generate a random number, fit a distribution, and then carry out a large number of simulations to find the mean and standard deviation of our Probability Generating Unit (PGU). By utilizing the mean and standard deviation of each input into the model, we can analyse the sensitivity of the model, which constitutes a global sensitivity analysis.



Moving forward, let's delve into machine learning in hydrology. Sensitivity analysis aside, let's now focus on mathematical models. We've covered model formulation, modelling steps, model calibration, validation, evaluation, and sensitivity analysis. Now, we shift our attention to machine learning in hydrology, a pertinent topic in today's landscape. Machine learning is a subset of artificial intelligence (AI), so it's apt to start with an overview of AI.

Artificial intelligence refers to the intelligence demonstrated by machines, distinguishing it from natural systems like humans or animals. The field of AI research revolves around studying intelligent agents, devices capable of perceiving their environment and taking actions to maximize the likelihood of achieving specific goals. In this domain, we train machines, computers, or any intelligent agent to learn and act towards predefined objectives. AI is characterized by machines emulating cognitive functions associated with human minds, such as learning and problem-solving.

And in the very beginning, in the first lecture, I talked about ah, learning. When I was talking about learning outcomes, ah, I mentioned that typically there are 6 levels of ah, cognitive skills as far as human beings are concerned, starting from ah, remembering or knowing to analysing, ah, and so on. So, that is shown by human beings, but typically ah, in artificial intelligence, we train the machines so that they are able to also show the same intelligence as the human beings, that is for learning or for problem-solving.

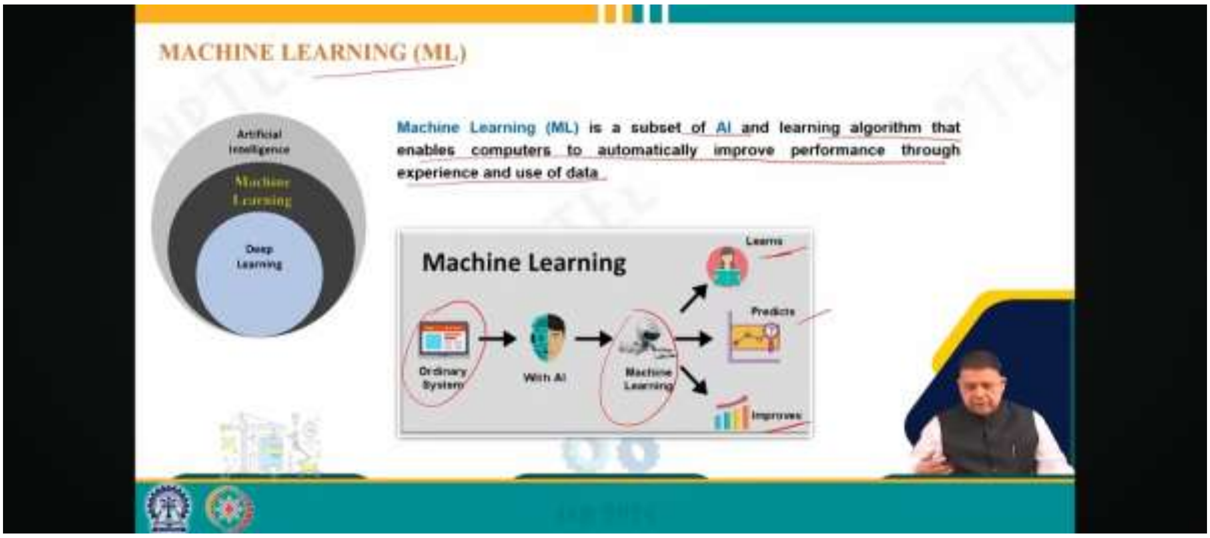


Then, as we know our day-to-day life is full of applications of artificial intelligence nowadays. For example, ah, you have Alexa or Billie or something, you just give instruction and then it can play music for you. Maps and navigations, almost everyday life, we use ah, these maps and navigations which work on artificial intelligence. Then, facial detection and recognition, ah, in our ah, smartphones, we always use. The same is true with biometric and phone touch ID. So, many of the colleges, you might have seen the attendance based on biometric ah, ah, some fingerprinting or something like that, or even the smartphones ah, use ah, the touch ID. Then, search and recommendation algorithms. So, you type in Google something and then you will get hundreds of possible recommendations. So, that is also based on artificial intelligence.

Ah, similarly, nowadays, a very common chat box, you talk to the machine, basically. So, based on the keywords you provide, the machine will answer. And then, of course, the social media feed and advertisements, you just look for a particular item, say shoe, and then any other page you open, you, you, you open a news , website or just any other website, and you will find advertisements related to that particular search item which you used. Other than that, in our day-to-day life, Netflix or Amazon Prime, ah, there also we find that they take into account our previous watches or history and then they give us recommendations.

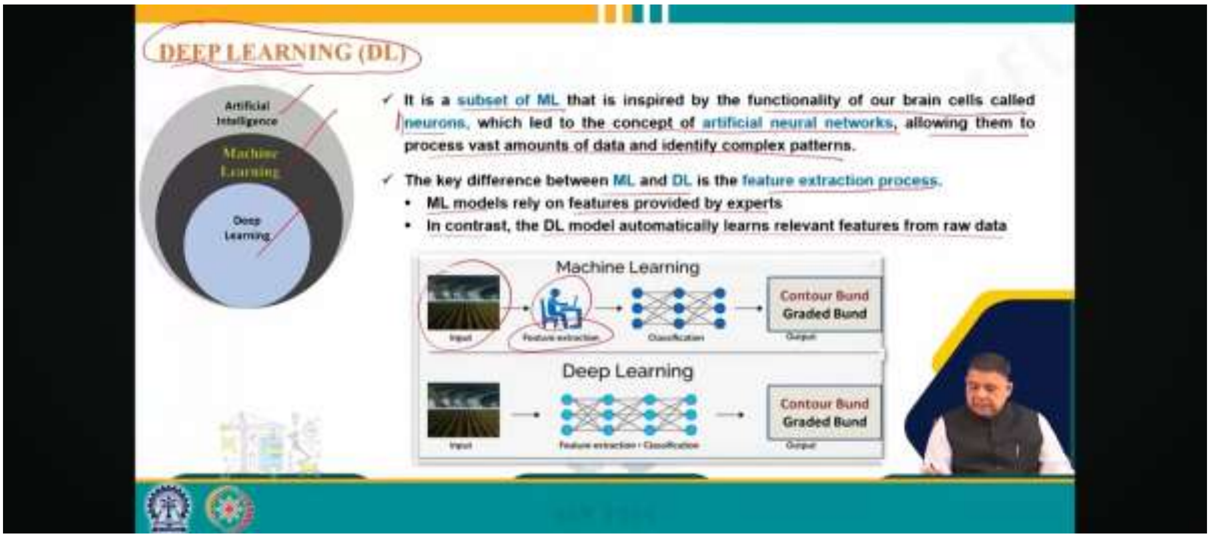
So, artificial intelligence has come in a big way and it is an integral part of our day-to-day life. So, that is why it is important to be with it and to learn about it, and that is how I am introducing machine learning in hydrology. As for machine learning, it is a subset of artificial intelligence and a learning algorithm that enables computers to automatically improve performance through experience and the use of data. So, of course, if you have an ordinary system and if we have artificial intelligence machine learning, then obviously, the system learns from the data which is provided to it. It learns it, of course, makes predictions and also as more and more data come, it improves its performance. So, an automatic learning procedure is there; it improves its own performance by learning through the data or data mining, we could say.

So, that is machine learning and of course, we have very commonly used nowadays deep learning, which is a subset of machine learning that is inspired by the functionality of our brain cells called neurons, which led to the concept of artificial neural networks allowing them to process a vast amount of data and identify complex patterns. Of course, for an artificial neural network, the application of such concept even in the early 90s, this technique came and we have been using artificial neural network, but nowadays, at the same time, the term which is commonly used is deep learning. So, we have artificial intelligence subset as machine learning, and a subset of machine learning is deep learning. And the key difference between machine learning and deep learning lies in the feature extraction process. Machine learning models rely on features provided by experts, whereas in contrast, the deep learning models automatically learn relevant features from raw data.



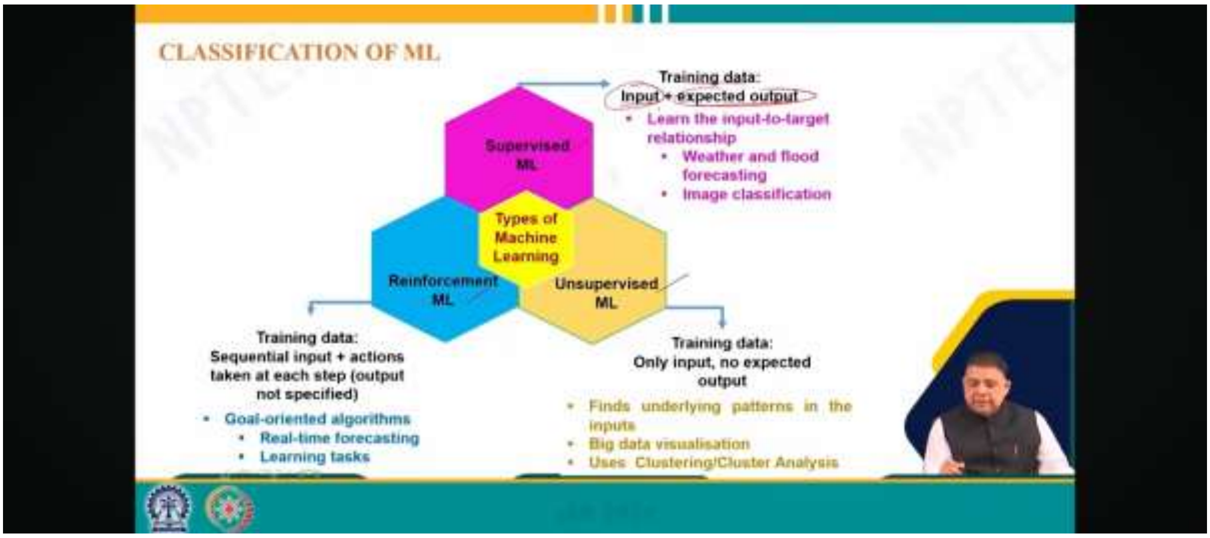
So, if you have a machine learning model, you have input, and then of course, ah the experts on expert will be needed to provide some feature extractions ah it will feed the classification and the output will be, say, contour bund or graded bund. Then if it is a deep learning model, then obviously, it will be provided input and it will ah do feature extraction and classification it will automatically do and come work with the same output. So, the role of ah the expert is ah taken care of automatically by deep learning. Coming to classification of ah machine learning a models, we have ah 3 different types of machine learning: supervised machine learning, unsupervised machine learning, and reinforcement machine learning. So, in supervised learning, we provide training data that is in the form of input and expected output.

So, we provide both input and expected output, and ah the model learns the input to target relationship, and it is very commonly used for weather and flood forecasting or for image classification. So, supervised learning. When it comes to unsupervised learning, we, as far as training data goes, we only provide input and no expected output is provided to the model. So, basically the model finds underlying patterns in the inputs through big data equalization and typically it uses clustering or cluster analysis to really come up with understanding or finding the underlying patterns. In the case of reinforcement machine learning the training data here is sequential input, that means, it is not ah static is supervised or unsupervised learning, it is more dynamic in nature that is we provide sequential input and actions taken at each step and output we do not specify here.

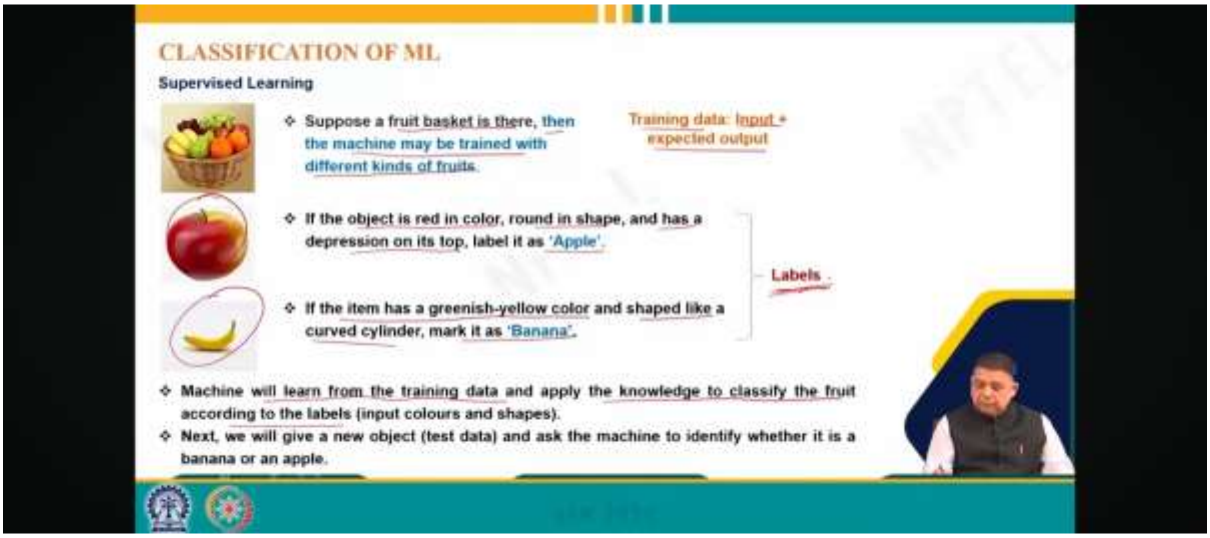


So, it is more similar to unsupervised machine learning from that perspective in which no output is provided, but it is more dynamic in nature as I already mentioned. And of course, you have goal-oriented algorithms which can be used for real-time forecasting or for learning tasks. So, robotics and all this is very commonly used. These are different machine learning techniques. Now, coming to supervised learning, suppose a food basket is there, then the machine may be trained with different kinds of fruits. So, obviously, training data is input and expected output both in this case we will provide.

So, if we train the model in the sense that if the object is red in colour, round in shape, and has a depression on its top like here, then it is an apple and then we will provide the label. Similarly, if the item has a greenish-yellow colour and shape like a curved cylinder, then it is the mark it is banana, so that means we provide a label. So, we provide the output and then we train it based on in this case based on the colour based on the shape and based on other particular shape features whether it is apple or banana or any other fruit for that matter. So, we provide labels and the machine will learn from the training data and apply the knowledge to classify the food according to the labels, that is input colours and shapes which we are providing here. So, next time if you give a new object, say test data and ask the machine to identify whether it is a banana or apple then based on these labels, based on the training, it will come up with the result that it is banana or apple or any other matter for that matter.

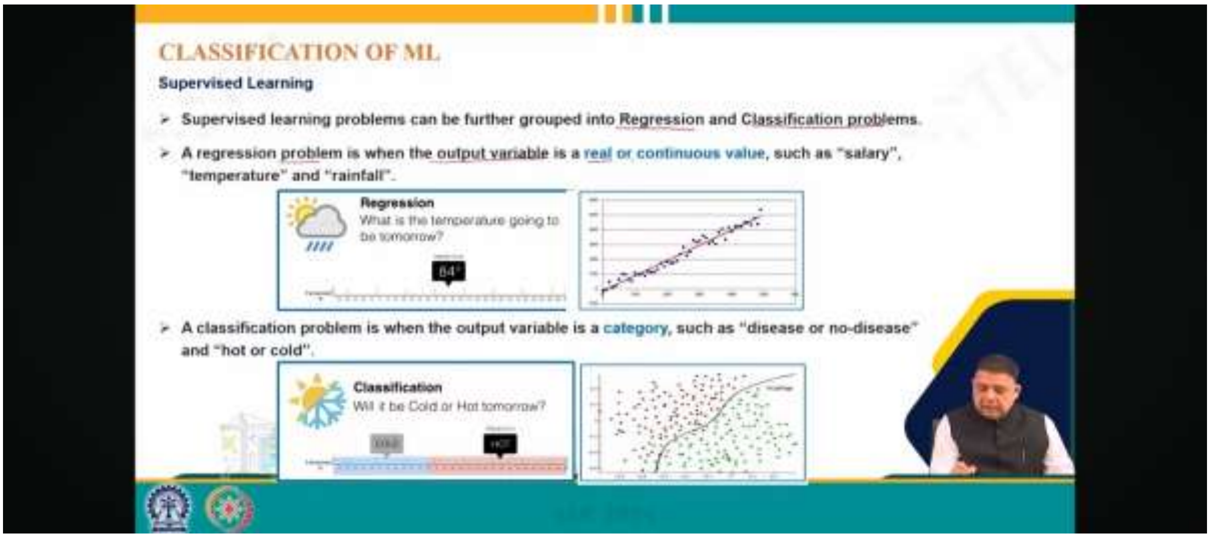


So, that is supervised learning, and supervising learning problems can be further grouped into either regression or classification types of problems. A regression problem is when the output variable is a real or continuous value, for example, salary, temperature, or rainfall. So, any variable that has a real or continuous value can be a regression problem and can be fitted for that. So, based on the data, what is the temperature going to be tomorrow, you will be getting the result because you have a fitted curve available. So, based on that the model will answer what the temperature will be tomorrow.

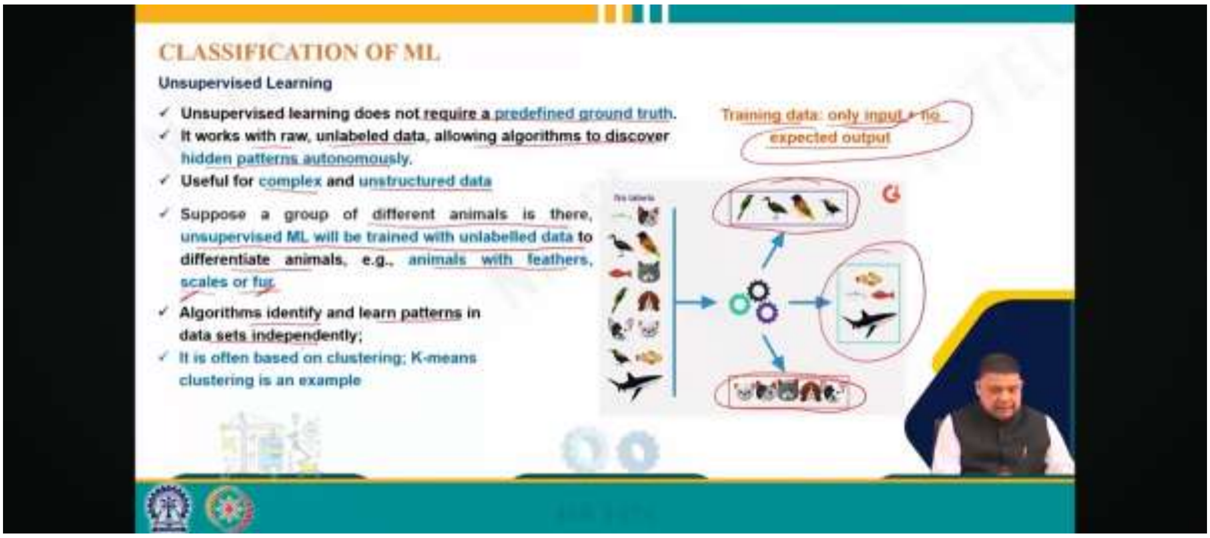


On the other hand, a classification problem is when the output variable is a category such as disease or no disease, hot and cold. So, obviously, based on this characteristic, data are segregated or classified. So, based on whether it will be cold or hot tomorrow, depending upon which condition a particular data belongs to, the model will tell whether it will be cold or hot, or similarly whether it will be a disease or no disease, or any other category for that matter. So, that is supervised learning. Then we come to unsupervised learning and in unsupervised learning, as we mentioned, we training data we only provide the input and no expected output is provided here, in fact and basically, unsupervised learning does not require a predefined ground to that means we do not provide the expected output data. It works with the raw, unlabelled data allowing algorithms to discover hidden patterns autonomously. So, it is more dependent on the data mining itself. So, within the data itself, we try to learn and it is useful for complex and unstructured data. So, for example, suppose a group of different animals is there, then unsupervised machine learning will be trained with unlabelled data to differentiate animals.

So, animals with feathers, animals with scales and animals with fur, for example. So, if we say animals with feathers, then obviously, the model will try to learn and segregate all birds together. If we talk about animals with scales, then again, the model will try to fathom through the data and try to segregate all fish-like animals in another group. And if we talk about animals with fur, then of course, the cats, the dogs, rabbits, and all animals which have fur, it will try to group into a different set. So, basically, based on the data itself, it will try to learn the hidden patterns autonomously and try to group the data.



So, algorithms identify and learn patterns in a dataset independently. It is often based on clustering, like k-means clustering for example, which is an example of an algorithm used for unsupervised learning. So, this is basically how we train in unsupervised learning. Then lastly comes reinforcement learning, where in the training data we give sequential input and action taken at each step. So, unlike unsupervised and supervised ML, reinforcement learning does not rely on a static dataset, but operates in a dynamic environment and learns from collective experience in the dynamic environment because there is sequential input and sequential instructions that need to be taken.



So, data points or experiences are collected during training to trial and error between the environment and a software agent say for example, a robot. So, based on rewarding desired behaviours and punishing negative behaviours, it functions. For example, let us say that we provide the agent with two conditions and we select a certain action policy. Say for example, we ask the agent to go look for a colder environment and then, of course, based on this policy, it takes some action and instead of going to a cold environment, it goes to fire.

So, obviously, because it has not followed the action properly. So, a penalty will be, say, a penalty of 50 points or whatever is prescribed, that will be put. And so, obviously, we get reward or penalty based on whether it has followed the instruction correctly or not. So, then obviously, it will learn that fire is bad and next time it will avoid it. So, that is the learning from this particular step and then of course, we will provide the next policy, and then we will iterate until an optimal policy is found. So, the agent learns to seek maximum overall reward to achieve an optimal solution. So, obviously, if it does the positive way that is looked for water.

So, if it finds water, it will be rewarded water in a bucket or water on the surface, then again it could be rewarded, whether even it followed. So, based on those sequential policy changes, it will learn and maximize the overall reward to get an optimal solution and this reinforcement learning is very popularly used in gaming and robotics. So, with this, we come to the end of this lecture, and we have started with the sensitivity analysis of hydrological models and then switched over to machine learning in hydrology. We have taken only a part of it. Next lecture, we will take forward this machine learning in hydrology. But we have learned that there is artificial intelligence, then machine learning and deep learning. And of course, within the machine learning, there are three different types of algorithms that are possible. Thank you very much. Please give your feedback and also raise your questions or doubts. We will be happy to answer.

Thank you very much.