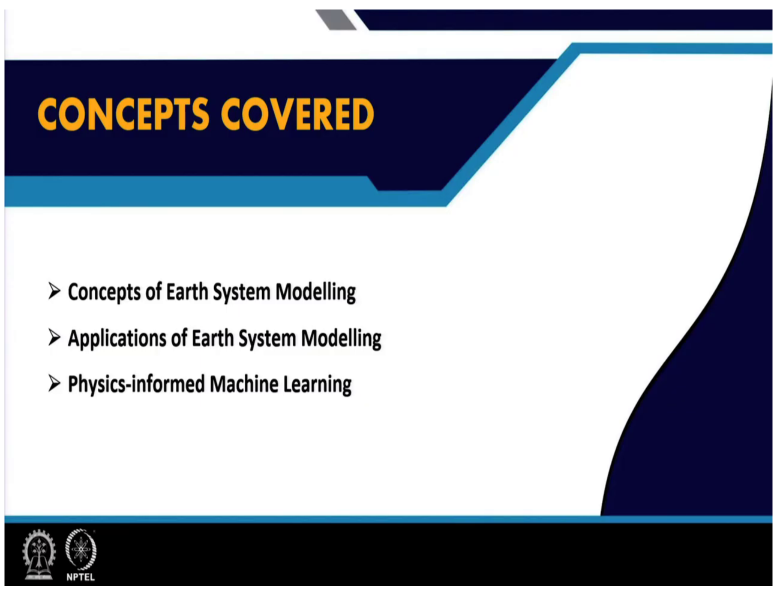
Machine Learning for Earth System Sciences
Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 05
Machine Learning for Earth System Modelling
Lecture - 34
Introduction to Earth System Modelling

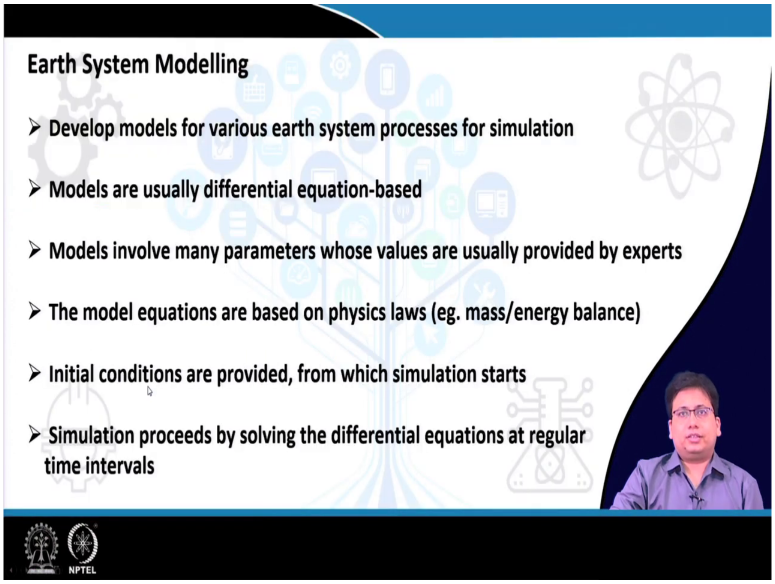
Hello everyone. Welcome to lecture 34 of this course on Machine Learning for Earth System Science. Today, we are beginning module 5, where we will be discussing the applications of Machine Learning for Earth System Modelling. The topic of this lecture is just to introduce you to earth system modelling and give an idea of how machine learning can help.

(Refer Slide Time: 00:46)



So, we will discuss various concepts of earth system modelling and we will see their applications also. And finally, we will also discuss this concept of physics-informed machine learning.

(Refer Slide Time: 00:59)



So, like this is just like the today we will not discuss any research papers, we will just like see the our understand the different concepts of earth system modelling. So, first of all what is earth system modelling? So, in different stages of this course, we have been making a references to different process-based models. We have talked about Numerical Weather Prediction of NWP or using weather research and forecast. We have talked about climate, global climate models, regional climate models and so on.

So, the like we have also mentioned that there are these process-based models for a different kinds of earth system processes like for climate, hydrology and all these things. So, what do these models do? They basically simulate the earth system processes and most of these models are based on differential equations, that is to say there is a set of equations which we call as the governing equations of the models.

So, like the there are of course, the variables corresponding to the actual meteorological variables and so on. And so many so, what these most of these process-based models do is that they define certain differential equations over these variables. Say for example, there are some of them are say toy models like Lorenz 96, Lorenz 96 is a toy model which shows which considers two variables. You can say a fast-moving variable like whose value changes rapidly and a slow-moving variable which whose values changes slowly.

But, the like the they have they are coupled, that is the slow moving variable has a has an impact on the fast moving variable and vice versa. So, like basically like the time derivatives of both variables are mentioned, let us on do using two differential equations. So, those two are the governing equations of the Lorenz 96 system.

So, that is our of course, a toy model as I said. But, the actual models for the different processes, they also have governing equations which like express the time derivatives or whatever of the different variables in terms of in terms of that is each other and so on. And the so, most of these model equations they are based on some laws of physics.

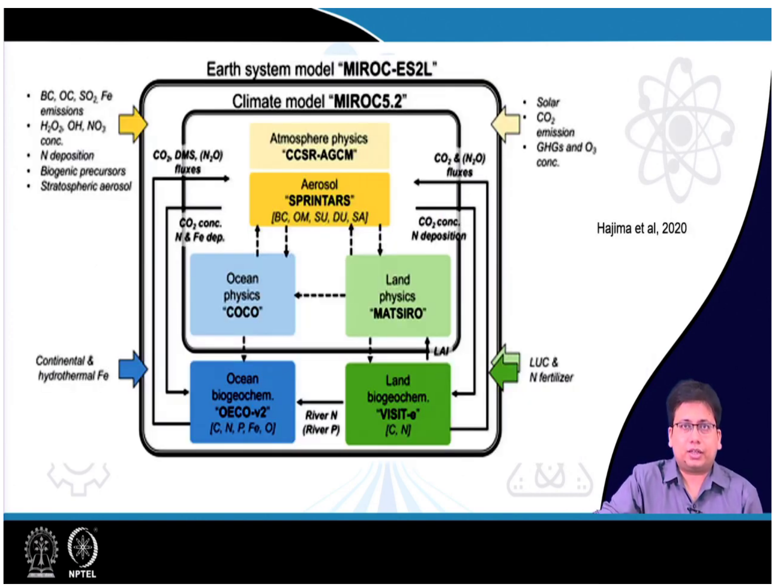
Say for example, conservation of mass, conservation of balance energy and so on. And, now the whenever we want to do any kind of simulation by these models, the initial conditions are provided from which the simulation starts. And, thus this simulation proceeds by solving the differential equations at regular time intervals.

(Refer Slide Time: 03:44)

So, like the US Energy Department has brought out this like this cartoon. So, it shows different sets or we can say different families of earth system models like. So, these in the for the atmosphere we have the global water model, extreme water, I mean extreme weather model, clouds and convection model, aerosol models and so on. Similarly, there for land we have the soil hydrology models or vegetation dynamics models, plant process models, human interaction models.

For ocean also we have ocean current and Eddie model, this salinity model, sea ice model. Similarly, for the cryosphere for the ice based models at like the glacier models, ice sheet models etcetera. So, like each of these themselves can be called as earth system models in their own right. But, a an integrated earth system model will have all these into like we will take all of these into account.

(Refer Slide Time: 04:47)



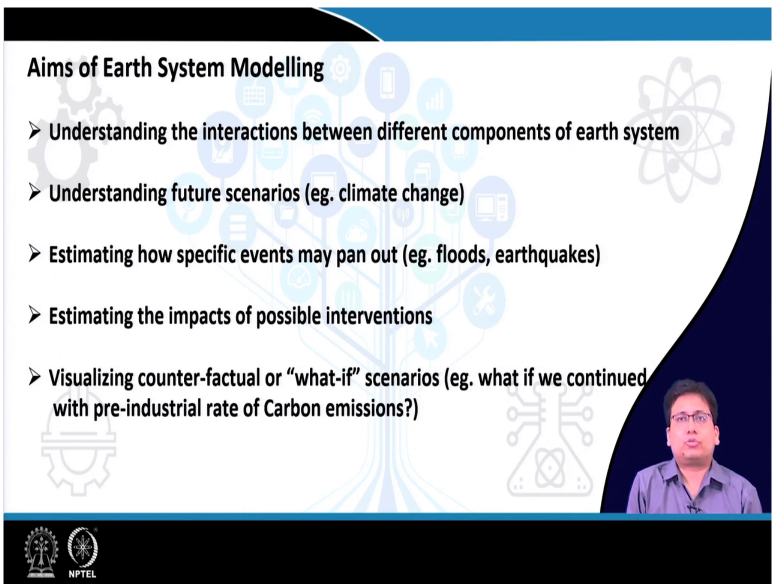
Say for example, this MIROC is ESL ES2L. So, like that is this MIROC is basically the name of an agency which has brought out this model. So, this is supposed to be a like a an integrated earth system model. It was initially developed as the as a climate model only. So, like which includes the atmospheric physics, the aeros, the ocean physics and the land physics.

And, but the earth system model or like retains this climate model, but also adds some other components like the ocean bio geochemistry, the land biogeochemistry and so on. It also the so, this model this earth system model is also receives the various extraneous variables. So, if you remember in some in like our earlier lectures where we are discussing how to develop statistical models using latent variables and so on.

We mentioned covariates or extraneous variables. So, like lots of these variables are intraneous to like intrinsic to the model, that is the model actually specifies equations for like for how there those variables may behave. But, then there are certain other variables or covariates which are not explicitly modeled.

So, they are provided as inputs like external inputs to the model. These include the say the solar emissions, the carbon dioxide emissions, greenhouse gases etcetera. So, these are either they come from outside the earth system as in case of solar or they come from human dynamics which are not of course, modeled in this taken care of in this model.

(Refer Slide Time: 06:38)



Now, what is the aim of earth system modelling? What do we hope to get out of it? So, first of all we want to understand the interactions between the different components of the earth system. So, like as you can see so, like here like we there are so, many different components. And, as you can see these arrows, these indicate the interactions between the different components. These MATSIRO or COCO as is written here, these are actually sub process model.

So, for oceanic physics there is this COCO model, for land physics there is this MATSIRO model. So, these now these in these arrows indicate that some outputs of these models are being provided as inputs to the ocean models. And, similarly here there is a bidirectional input output relation between the ocean model and the aerosol model, that is, the aerosol impacts the ocean and the ocean also impacts the aerosol and so on and so forth.

So, these kinds of interactions these like we can like we can form a better idea of these like interactions between the different components using such models. Then, like arguably the most main reason why these models are so important now is for understanding the future scenarios. So, say for example, climate change, that is these models might be as the these General Circulation Models, the GCMs has developed by various the various research groups under the IPCC, the Intergovernmental Panel for Climate Change.

So, these models are run for several years or several decades into the future and they give a some idea about how the world's climate might be in 2050 or 2080 or something like that. Now of course, that that depends on certain extraneous variables like they say how much carbon will be emitted by the human beings depending on the result of their of the amount of industrialization we achieve or urbanization we achieve or whatever technologies we adapt and so on.

So, those are the extraneous variables to the model. But, what it does is under different scenarios related to emissions; these models try to give us like an idea about what the future climate may look like. And, also like the like these that is of course, a very far into the future, but what about something very specific or a specific event, how a particular event may pan out. Let say for example, a flood or a cyclone or an earthquake. Say, suppose a cyclone originate somewhere in the Bay of Bengal or at this particular location in the Bay of Bengal.

Then, in what ways it can move, which all regions it can affect, at what speed it may make the land fall and how much damage it can cause? So, these kinds or the or what or what is the specific or how much time does it, we will we can it possibly take to make a landfall? So, also like we can understand a how a specific critical event like that may pan out; so, that we can have some idea about how to plan for such an emergency.

Then, we can also estimate the impact of possible interventions. So, suppose like we decide to scale down on industrialization or suppose we start using more of solar technology by eliminating the use of fossil fuels, how does that alter the climate change scenario?

So, or like or even for a very more localized process, if I make some kind of interventions so, like for example, like here also as you can see these extraneous variables. So, suppose I artificially change some of these extraneous variables, how does the result change as a result of that?

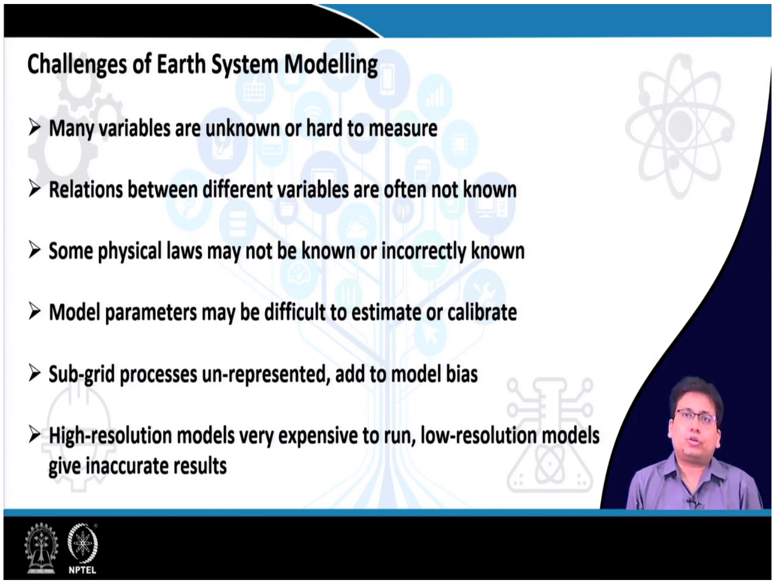
And finally, visualizing various counter factual scenarios. So, counter factual scenario means something that has like is a hypothetical scenario which has not happened and which we have no idea no way of knowing. Say for example, now we every year we see several cyclonic storms happening in the Bay of Bengal which strike the eastern coast of India. Several way like two, three of them are happening every year; earlier also they used to happen, but it seems the frequency has increased, their intensity may also have increased.

Now, some people say that this is happening as a result of climate change. Some other people may say that there is no climate change; otherwise also it would have happened. So, like then the question is had climate change not happened that is like say suppose we had still been stuck with the climate as it was say 150 or 200 years earlier, then also could we have expected this kind of an event, that there are being an increase of frequency of cyclones and so on; could we have expected such a thing?

So, there is no prima facie, there is no answer to this question, because we because the climate has changed. It is not what it was say 150 or 200 years ago. It is then how can we possibly say what would have happened if the climate change had not happened. So, these kinds of what if analysis, they can be achieved with the help of these models. It is possible to initialize the model to the pre-industrial period say around 1840 and so on.

And, then without any further carbon forcing just allowed the model to run till this time and, then we could see if actually there are certain periods where these cyclones over Bay of Bengal had become more frequent or not.

(Refer Slide Time: 12:17)



Now, this process of developing earth system models, this is fraught with challenges. So, challenges are manifold. So, first of all the as I said these models have governing equations which are like which indicate how the different variables interact with each other. The problem is that, first of all many variables are unknown, that is we do not even know which all variables are may be crucial or may be useful for these processes.

And, even if we know like we may not be able to measure them. So, there are certain there may be certain variables, say we know that there is wind speed 500 meter or 5 kilometers above the earth surface and so on. But, such the variables are difficult to measure. So, you like unless we have measurements, there is like there is no way to validate the model or to parameterize the model and so on.

Then secondly, the relations between the different variables are also often not known, also the governing equations of the models which I we are talking about. So, these are based on certain physics based laws, but those laws have been devised by scientists based on empirical observations. Those laws may not necessarily be perfectly true or they may or they may be only approximately true, but certain other factors may have to be brought into it.

Say for example, Newton's laws of gravitation or like say for example, if I release a ball from a high point and allow it to fall to the earth surface. Now, Newton's laws of course, tell me how long it will take to hit the earth surface and so on. But, in addition there are many other effects like say air resistance, viscosity and so on which are not taken care of by Newton's laws of motion. So, those are like additional corrections which we may have to make.

So, similarly in these governing equations also, like we it is it might be necessary to make various other corrections bringing more effects and so on which we may not know as of now. Then finally, these models or the governing equations, they have various parameters, coefficients and so on, which whose actual values we may not know. Now, these so, so first of all those good things may be entirely hypothetical and secondly, like their values may change from location to location, from time to time etcetera.

Now, these usually what happens is that some experts specify the values of these parameters based on again some empirical observations. But, as I said the variations of such values and so on are usually like not taken care of in these models and they result in certain biases. Then, another very important thing is that every model has a particular resolution. So, we have been discussing right from the beginning about spatiotemporal modelling.

So, that is any model be they the these process models which we are talking about now for say or the statistical models which we had talked about in the early lectures, each of them will need a particular spatiotemporal structure to work grid system to work with. So, there will be certain special resolution, temporal resolution and so on. Now, whatever resolution we are working at let us say the $50\ km\ X\ 50\ km$ or something like that or may be like hourly or daily in temporal dimension.

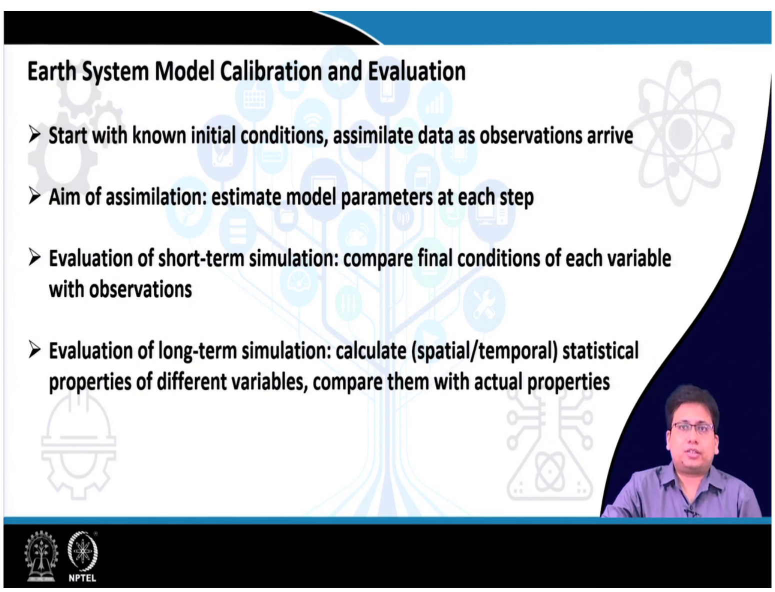
Now, there may be sub grid processes which we can also call as unresolved processes that is some process which is limited to say only 20 kilometers or say like which occurs over only a few seconds or a few minutes and so on. So, such processes are like they are not taken care, they are too small for the model to take care of. So, we call these as unresolved processes.

So, in the models we just like consider these sub grid processes as a source of error or at most maybe a some kind of parameter or coefficient right, the. So, like the problem is that like the various, but these processes are neither like a coefficient or nor are they some random noise. There is actually a process behind it, only we are not able to take it into account.

So, as a result of this like a some bias is often added in these models and which and these bias keeps getting accumulated. As a result of which as the process runs for the simulation runs for a long time the, like the output keeps further deviating, further and further from reality. And, most another very crucial factor is that the high resolution models like they are very expensive to run, that is one way to circumvent the problem of sub grid processes is to of course, reduce the grid size and make the model as high resolution as possible.

So, that there are lesser and lesser unresolved processes, but that also makes the model more much more like computationally challenging. And, a the typical climate model today requires like an elaborate super computing system to run. And, then also it runs for like a very long time and in the process it consumes a lot of electricity and emits carbon and so on. And, it may itself this process itself may be contributing to climate change. And so, in general this is a very big limitation of these kinds of models.

(Refer Slide Time: 18:13)



Also now ok; so, suppose we have the theoretical basis of a model somehow developed in spite of all these problems, then also first how do we calibrate the model and how do we evaluate the simulations? So, like one way to do it is like through with the help of data assimilation. So, like the model will start with certain initial conditions which we may have specified and then we will allow it to run.

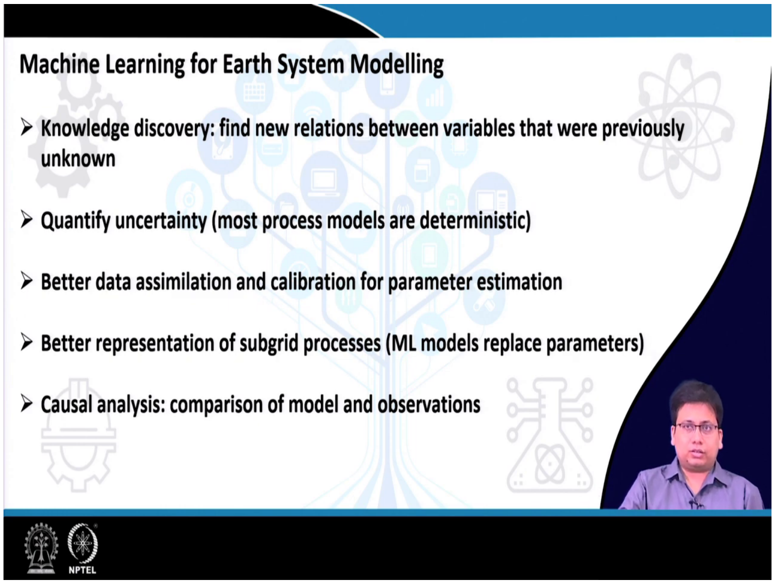
But, like as the model is running, we will keep providing it with observations that is the model will be estimating the values of different variables. But, we will be from time to time providing the actual values or observed values of those variables. So, that the model can calibrate itself and tune it is all these parameters in the at each step. So, this is basically the data assimilation process for which we have already studied one approach, the Kalman filters.

So, apart from the so, Kalman filter is only one way, there may be various other approaches to data assimilations which these models can be making use of. Then, if we are evaluating say short term simulations, then we can like compare the final output of each variable with the observations. But, if you are doing a long term simulation, then like its evaluation it is like it is kind of pointless to that is do a point wise comparison of the different variable values.

That, that is to say let us say I am simulating the Indian climate for say 50 years or 100 years or something like that. In that case, it will not make much sense to like that is every year see if the models that is simulated values match the corresponding years observations and so on. Because, like if we are running it for such a long period, then making a point to point measurement is like we cannot expect it to be accurate point by point. So, instead what we can do is like we can measure the various statistical properties.

Let us say we focus on one particular region, let us say Kharagpur and see whether the mean temperature of Kharagpur as simulated by the model over this period. Like is it does it resemble the actual mean temperature of Kharagpur for the same model as observed in reality or what about the variation, what about the maximum and minimum values and so on and so forth. So, this way like we do a statistical comparison of the simulated variables versus the observed variables.

(Refer Slide Time: 20:55)



Now so, this is all about earth system modelling. But, where it does machine learning come in here or in other words what can machine learning do as to help this process of earth system modelling? So, there are many things it can do. First of all knowledge discovery, as already mentioned the governing there are various governing equations used for these models. And, these are based on the our the laws of physics that we know.

But, then there may be some additional relations or additional laws, additional equations which we do not know. So, it is actually possible to use machine learning to find new equations. In fact, there is a paradigm of equation discovery which is used in like in physics-informed neural networks which we may briefly touch upon in one of the subsequent lectures in this module.

Then, most importantly uncertainty quantification that is most of the process models are deterministic. But the like, but we know that there is a like whenever we are making any kind of forecast or any kind of simulation into the future, we can never say anything in like in concrete terms. There will always have to be a notion of uncertainty that is instead of making a point wise prediction, we should always be able to give some kind of a range or confidence interval.

Now, most of the deterministic climate or earth system models are unable to do that to. So, to achieve some kind of or to quantify the uncertainty of these processes, one way to do it is through is by using machine learning, especially the probabilistic models. Next, better data assimilation and calibration for the parameter estimation so, we like machine learning already provides an elaborate framework for parameter estimation.

So, some of the like as I said earlier most of the parameters of the process earth system process models, they are like just expert suggested values which may have flaws. They may be varying over space, varying over time and so on which is like which no expert can specify beforehand.

But, using data like and also using data different frameworks for a parameter estimation which is known in to the machine learning community, it might be it might be possible to estimate these parameters in a much more robust way. And, then we can have a better representation of the sub grid processes which we just mentioned.

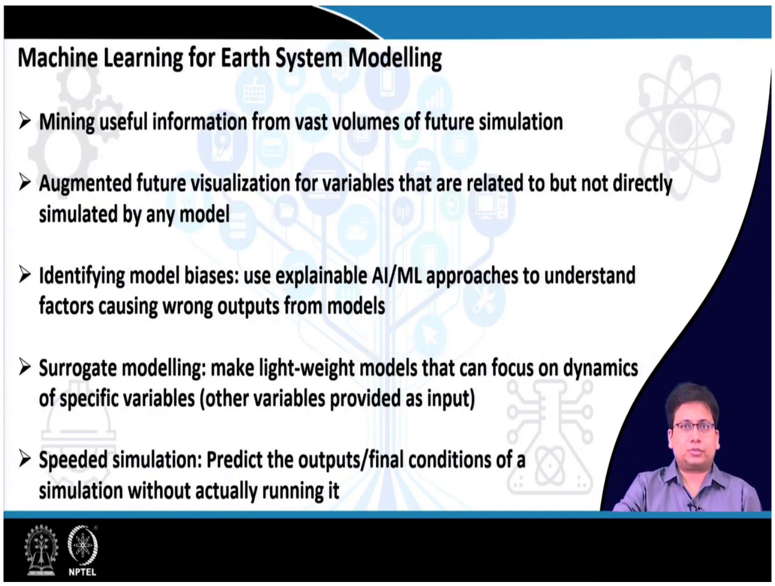
So, instead of just specifying as a random noise or some kind of like parameter or coefficient, the machine like we can instead have a machine learning model for the for those sub grid processes that is the like the earth system model can work in tandem with machine learning algorithms which like which you can say are like proxies for those sub grid processes. That is they do not actually simulate those sub grid processes, but like given a certain condition like the machine learning model can actually predict what like what those sub grid processes will cause.

And, then the those causes of those sub grid processes, I am sorry the effects of those sub grid processes can be then plugged into the earth system model for further simulations. So, this is one very important way in which machine learning can replace the or like can help in the better representation of the sub grid processes which is a proverbial Achilles heels for the earth system models. And, then causal analysis we have already talked about, that is like an earth system model is not very useful unless it can actually preserve all the causal relations.

One reason for this is we have mentioned one of the most important applications of earth system model is in understanding how the system may react to some kind of intervention. That is, if I manually like a reduce the carbon dioxide content in the atmosphere by some kind of manual intervention, then what will happen to the system? So, like for that so, this is a question of causality. So, carbon dioxide is the cause, but what are the different effects?

So, like unless the earth system model that we are using is able to like preserve this causal relations, then it is not very useful that is then it will not be able to answer the question which I just asked in a very accurate way. So, it is necessary that the any earth system model should be able to preserve all the causal relations that are actually observed in the like from real observations.

(Refer Slide Time: 26:00)



So, this is like we have already discussed the concept of causality and how machine learning can be used for causal inference and so on. So, this is another way in which machine learning can help. Then finally, this when these earth system models are run for a very long time we will so, they will generate a huge volume of data. Now, from we from that how can we extract useful information?

Say, suppose I want to understand say how many droughts will be taking place in such and such scenario of climate change or how many rainfall annual events will take place. So, for that we need various machine learning algorithms, because we cannot expect to sit and manually analyze those vast volumes of simulation data. Then, they I like I may also be interested in certain variables which are not directly modeled by the or I mean which are not directly simulated by the model.

But I may, but what they may be related to the variables which are actually simulated so, in that case that relation between the simulated variable and the target variable that may be represented by a machine learning framework. And then so, based on the various future scenarios or various simulated scenarios, I may be able to know the possible values of the target variable that I am interested in.

Even though the process model, the earth system model itself does not give me that. Then, apart from that like we can we have already discussed explainable and interpretable machine learning techniques, those layer wise relevance propagation and backward optimization and things like that. So, suppose a model makes a prediction, the idea is to understand why it made such a prediction that is which part of the input caused it to make such a prediction.

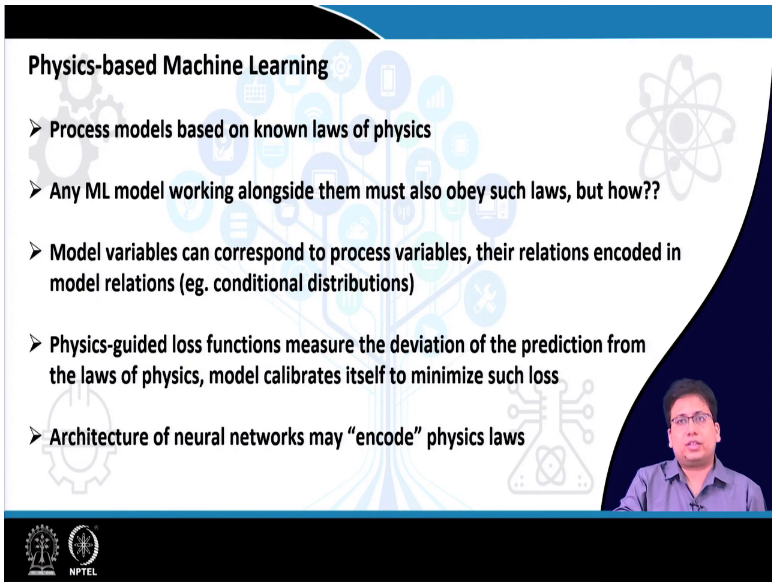
So, suppose an earth system model actually gives me some wrong result that is I am I simulate it for a particular duration and I from starting from an initial condition. And, I find that what it gives me as output does not match what is actually observed. So, then I want to understand what went wrong, which components of it gave the wrong results. So, like so, then we one way to do it is to use some of these like explainable AI techniques to see which parts of the input or the initial conditions are actually result important for determining the output.

And, that will give me some idea about like what is wrong with the model, is it that it is not taking certain parts of the input into consideration or is it putting extra emphasis on some other parts and so on and so forth. Then, one very important way in which machine learning can help is in developing surrogate models. So, what are these surrogate models? So, we have already discussed that these the process earth system models, they require very high computational power and time.

So, now, instead of running those huge models, can we actually use some low cost machine learning models, which does not aim to simulate the entire earth system, but maybe focusing only on certain parts of the earth system. And, like it will so, for a for any earth system model, it will be able the machine learning model will be able to tell me what that earth system model will simulate for any given input for a particular variable.

So, like so, it in a sense it is like learning the behavior of the model, given any particular input it should be my machine learning model will be able to tell what the process model is going to give me. So, this helps in speeded simulations also.

(Refer Slide Time: 30:03)



Now, how is this achieved? One way in which this kind the like machine learning can be used for all these things in earth system models is with the help of physics based machine learning. That is the machine learning we will not like of course, the all the machine learning algorithms that we are familiar with the neural networks, the random forest, the Bayesian networks, Markov random fields etcetera.

They will all be there, but they will have to be tweaked in such a way that they are suitable for this purpose, that is they are consistent with the physics of the system. So, the process models are all based on the laws of physics. But, the machine learning models as such are not based on that, that is it is possible to for a machine learning model to make such a prediction which will not obey any law of physics. Say for example, conservation of law, conservation of mass may be violated or something like that.

So, one way to achieve that is like is first of all through the model variables. So, like some of the model variables of the machine learning model, they can corresponding correspond to the process variables. So, like if you consider if you remember the statistical models which we discussed in the in our early lectures.

So, some of those the latent variables, they are supposed to actually represent the I mean latent and observed variables, some of them are supposed to represent actual physical variables. And, the natures of those variables, the typical range of values taken by the variables etcetera; can be like incorporated into the model variables also through things like say conditional probability distribution etcetera.

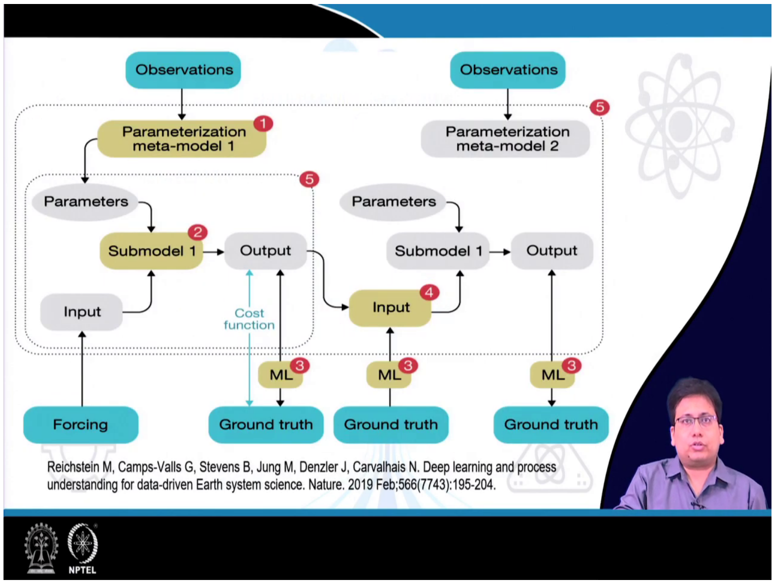
Then, we can like a very important way is to use the physics guided loss functions which in a sense measure the deviation of the prediction by the model with the laws of physics. So, let us say the machine learning model produces input y, corresponding to produces output y corresponding to input x.

So, but it turns out that the output y like is inconsistent with the physics, that is let us say conservation of mass is violated. In that case, we like we can design a loss function which shows by how much the mass has been I mean the mass before and the mass after, what is the difference between them.

So, ideally of course, it has to be 0, but like it will take some non-zero values. So, that non-zero value will be like recorded as the loss function. And, then we know when we are training the model, the model parameters are optimized so, as to minimize the loss function. So, it will try to like tune itself, calibrate itself in such a way that the this physics based laws will be nearly equal to 0, that is the outputs will have to follow the that particular law of physics.

So, like if we have if there is such a model which whose output must follow multiple laws of physics. So, for each of them we can have a separate loss function and all these loss functions can be added together; so, that all of them must simultaneously be minimized. And, then additionally it can also happen that the architecture of the neural network or for the architecture of the Bayesian network or whatever we are talking about, that itself encodes the various laws of physics in some way or the other.

(Refer Slide Time: 33:31)



So, this is the like this is a famous paper which appeared in 2019 which discussed how like Deep learning and process understanding for data driven Earth system science. So, like here like so, let us say these are the this is a sub model of an earth system model and a it has its own input as well as its own parameters and it has outputs and so on. And, then there they are these observations and the ground truth labels etcetera.

So, now according to this cartoon, the machine learning model can be used as an interface between the ground truth and the model output that is the machine learning what the machine learning model will do is, it will compare the models outputs with the ground truth.

And send feedback to the model, that you need to calibrate yourself to bring down the laws and so on. And, the machine learning, alternatively the machine learning model can itself like indicate or be like used to say estimate the various values of the model parameters for which the models output will estimate the ground truth and so on.

(Refer Slide Time: 34:45)

So, these are the different or a few relevant papers related to how to some of them give a examples of earth system models and some of them give some idea about how machine learning can help in these earth system models. So, like even a 3, 4 years earlier the earth scientists were not particularly keen on this matter.

But, nowadays for the last few years, our realization has been coming to the climate to the earth system scientist, that machine learning can actually can not only help, but perhaps even like take over the way they have been doing the earth system models. And, that is why we are hearing of terms like neural earth system models, where the model equations will no longer be based on say differential equations as they had been earlier.

But by machine learning models on neural networks or Bayesian or graphical models themselves; so, in the coming lectures, the last few lectures of this course, we will see certain applications of these earth system models for various specific applications.

So, till then bye.