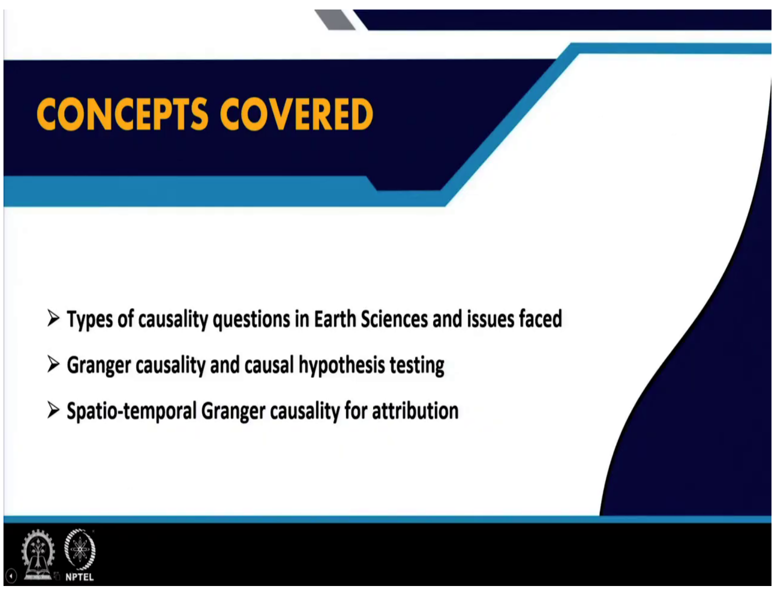
Machine Learning for Earth System Sciences
Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 03
Machine Learning for Discovering New Insights
Lecture - 19
Identifying Causal Relations from Time-Series - 1

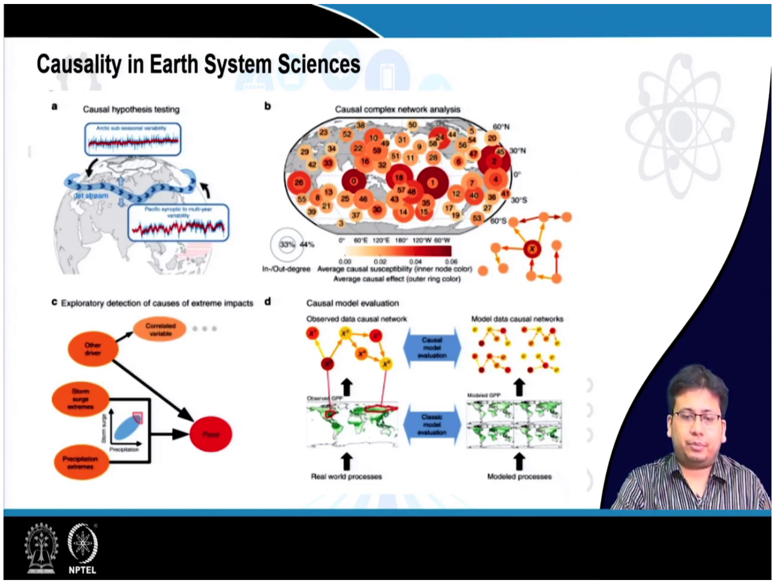
Hello everyone. Welcome to lecture 19 of this course on Machine Learning for Earth System Science. We are in module 3 which is where we are using machine learning for discovering new insights related to the earth system. So, in the topic of this lecture is going to be identifying causal relations from time series. So, this topic we will cover in two lectures. This is the first of those two.

(Refer Slide Time: 00:52)



So, the topic of causality we have discussed briefly in the in module 1, in these lectures we are going to see some concrete use cases. The concepts we are going to cover in this lecture are, first of all we will like recapitulate the different causality questions in earth sciences and the issues we face while solving them. The secondly, we will see some applications of Granger causality and causal hypothesis testing. And then, we will also see spatio-temporal Granger causality for attribution.

(Refer Slide Time: 01:24)



So, like here is like some the broad questions which or related to causality in earth system sciences. We had briefly touched over this in one of the lectures of module 1. Today, we will see, in this lecture and in the following lecture, we will see some use cases of each of them. So, the first type of use case is causal hypothesis testing. This is like, here we make may make certain kinds of hypothesis regarding a cause-effect relation between let us say different earth or geoscientific phenomena in different parts of the world.

So, climate scientists or earth scientists have various understandings or various notions about like which event or which kind of variable influences which other variable and so on. But these are mostly hypothesis. For example, it is known from empirical data that the El Nino-Southern Oscillation which happens in the central Pacific Ocean has a bearing on rainfall in India, monsoon rainfall in India. So, like it, one hypothesis there is that the like El Nino or that ENSO phenomena is a cause and the Indian monsoon is the effect.

But how far is this true? I mean correlation is there, but as we have discussed earlier also correlation and causation are not the same. So, here the hypothesis is that like one is the cause and the other is the effect, but it could also be that there it is the Indian monsoon which is the cause and the El Nino-Southern Oscillation is the effect. Or it could even be that there is a bidirectional relation. Or it might even be that like neither of this is the cause or the effect, but there is some confounder which is the common cause of all of these. So, which of these possibilities is true?

So, like here we can make a hypothesis. Now, we need to test the hypothesis. So, you might be familiar with the idea of hypothesis testing which is an important question in the domain of statistics. So, we know that like we we start with a null hypothesis and we defined some a test statistic and we defined some criteria by which according to which the test, I mean the null hypothesis can either be accepted or rejected at a certain level of confidence and so on.

So, like that kind of hypothesis testing related to causality is something, is a question that frequently pops up in this domain. But then again there are sub questions like what should the test statistic be or like how exactly we can like evaluate the or the or test the hypothesis etcetera. So, these are like that will depend on the exact context in which the question is being asked.

The second set of problems is like developing something like a causal network all over the world. So, like climate networks and so on, like the networks based on correlations, then networks based on even synchronization etcetera we have discussed earlier. So, like where we like look upon the whole earth surface as something as a graph and every location or a every grid as a node and we put edges in between them.

Now, the that kind of a network can be a causal network where we can like draw an edge from a node u to node v, if we can be sure that node u is the cause, node u or as or a particular variable at node u is the cause of another variable at node v. So, that like that way we can develop some kind of a causal network. Of course, that will require us to do the causal analysis and identify lots of cause-effect pairs in different parts of the world.

So, one of the important or we like one of the most ambitious research questions in this domain is to build some kind of a causal complex network of the world. That will like enable us especially to understand that the phenomena of teleconnections in which like one phenomena in one part of the world is known to be the in resulting in another phenomena in another part of the world.

But then, like for that kind of relation to be understood in the cause-effect scenario, we need to understand the mechanism, the exact physical mechanism in which one very like one of them influences the other. So, that kind of thing can be achieved if we can build something like a causal complex network of the world. Then, another important set of problems is that of attribution, and especially the attributions of extreme events.

Say, for example, like a heavy rainfall event happened or a cyclone happened in some region, then can we say, can we attribute that thing to like a larger phenomena, like say climate change or something like that? Or can we say that had climate change not happened, had we like been at the same climate as let us say 200 years earlier then this particular event would not have happened. So, that is in general a very difficult question to answer because for that we need the counterfactual scenario. What if there was no climate change?

But then we of course, we do not know that because there is we have only one version of the reality. We do not know any other version of the reality in which there is no climate change. So then, how to answer these questions? So, this is the attribution problem which allows us or which aims to actually estimate what would have happened in counterfactual scenarios. So, since we do not have the counterfactual scenario, we have to somehow estimate its statistics based on whatever data we have available.

And then, see if a particular event of interest it like which actually happened we have to try to estimate whether it could have happened in the counterfactual scenario also. So, this kind of analysis, this is known as the at the attribution causes. The second one is or the sorry the fourth one is that of model evaluation. So, we know that or we have already discussed several times earlier in the previous lecture that there are these process models which are like which aim to simulate various earth system processes.

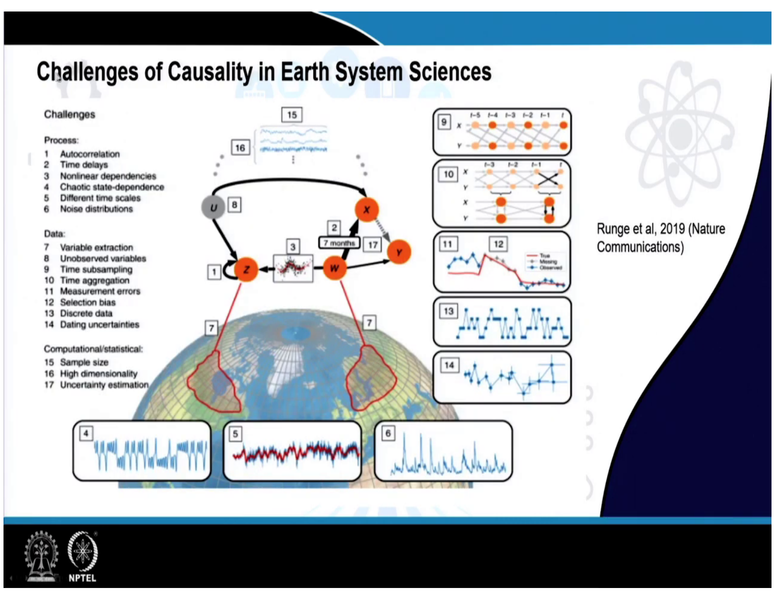
And the like, there might be hydrological models, there might be climate models, there might be ocean flow models, seismic models and so on. Now, the question is how realistic are those models. That is, like suppose there is indeed a causal relation from one variable to another in the in reality. That is we know that earth like geoscientific variable a influence at location s influences geoscientific variable b at location s'.

Then, is that kind of a causal relation is it preserved in the model simulations or like or there is merely some kind of simple correlation between the two. This is important because like these models these are used to evaluate the various counterfactual scenarios which we are which we just discussed. That is one possibility of evaluating a counterfactual scenario is to simulate it using the models.

But if the models they are not able to capture these causal relationships, then the like the simulation of these counterfactual scenarios that will be futile. Because, we will not know whether a particular the event which we are interested in would have happened in the counterfactual scenario or not because we do not know, because the model is not able to like understand its causes.

So, that it is whenever we build a simulation model, a process simulation model, it is very important to make sure that it is able to like preserve all the causal relationships that are present in reality. So, this is one important criteria according to which these models may be evaluated.

(Refer Slide Time: 10:14)



So, like now let us discuss some of the challenges that we face when we are applying causal analysis in the earth system sciences. So, like this we had like discussed a briefly in an another episode in module 1, let us discuss this in a bit greater detail now. So, the first challenge is autocorrelation. So, we know that like autocorrelation is when, like basically the values of one variable depends on the past values of the same variable.

So, like we if you remember the notion of Granger causality is aimed to like express or $Z(t)$ as a function of $Z(t-1)$, $Z(t-2)$ etcetera along with the potential cause which may we may cause a call as W. So, we can like in Granger causality we try to express Z, in the present value of Z in terms of the past values of Z as well as the past values of the potential causal variable.

Now, the if the autocorrelation is so strong, that is if the coefficients of the past values of Z tend to dominate the coefficients of the past values of the causal variable W, then even like even if some causal effect is there, like it the autocorrelation may be so strong that it cannot be separately discerned from the autocorrelation. So, that is like the presence of strong autocorrelation in the time series, hampers the causal analysis. Then, there is a question of time delays or time lags.

So, we have earlier also discussed that like especially when there is some kind of a tele-connection relation, say for example, these two phenomena which are happening in different parts of the world which are quite far apart. So, typically there is a time delay between any event that happens here and another event that happens here. So, like it might be difficult to understand how much time delay it is.

And if the time delay is quite long then it is possible that like in whatever causal analysis we are doing, the time delay will actually be beyond the window or beyond the time window which we are considering, so that the delayed effect will simply not be captured. Then there are non-linear dependencies. So, like in case of Granger causality, we had discussed that we try to express the future values of a variable in terms of the past values or as a linear function of the past values.

So, we use fit some kind of a regression model, but the relation can be non-linear also. Now, as I said the scientists are currently trying to solve this issue by using non-linear Granger causality. Then the comes the question of chaotic state dependence. So, like especially a chaotic time series is notoriously difficult to predict, because the nature of chaos is such that if there is some small deviation at any point of time, the future values may take a completely different trajectory.

So, in the presence of such chaos, identifying causal impacts is becomes very difficult. It might also happen that the different phenomena which we are studying happen at different time scales, that is one may be fast and another may be slow. So, if that is the case then I again identifying the like the causal impacts might become very difficult because like the causal variable that might be very slow moving and the observed variable or the effect variable that might be fast moving.

So, like we will find it very difficult to a like express one times, the impact of one time series on the other.

And apart from that there are also some challenges which arise out of the data. Say for example, like the in case of, there might be unobserved variables. Like, so these are like basically the confounders which we talked about. So, remember the example of shark attacks and ice cream sales in that case there was an unobserved variable the temperature.

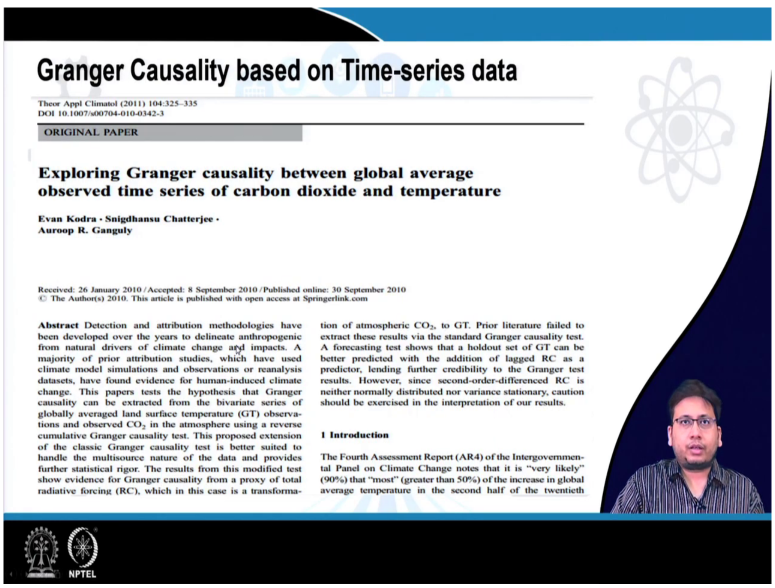
Now, similarly when we are trying to express one or try to understand the causality relation between two variable, it is possible that both of them are impacted by a third variable which we have not taken into account, which we are not aware of. So, in the or it may also be that that variable we may be aware of, but we have no way of measuring it. So, the presence of such unobserved variables makes the job even more difficult.

Then there are other issues like time subsampling and so on. This is related to the problem of different time scales. So, like we of course, cannot like measure that entire time series. We have to like any sensing device it will measure a particular variable at fixed intervals. Now, it might happen that the process is so far that changes take place within that or within that interval or rather it is the impact might be at a higher resolution than what we can sample it.

So, if that happens then we may not be able to identify the relations or may we may identify yet, write it as incorrectly. That is say for example, there might be a lag relation like this, but we may end up interpreting it as some kind of a contemporaneous relation. Then there are of course, problems related to measurement errors that is when we fail to measure the variables accurately and thereby make thereby we may end up coming up with relations that are spurious. Then, there can be selection bias where we focus only on one part of the data which is not representative of the entire data.

In present, if the data is discrete, then it becomes even more harder to because if you observe the Granger causality and such models they are all continuous valued distributions. But in case we are dealing with the discrete data, it will it becomes a more complicated situation. And then there are other kinds of uncertainties related to dating of the variables and other computational issues like the sample size, the dimensionality, the uncertainty and so on.

(Refer Slide Time: 16:55)



So, now let us come to a few use cases. So, in this, in the rest of this lecture and in the following lecture also we will see some use cases of causality in earth sciences. The first one is based on Granger causality of time series. So, here the task is to explore the Granger causality between global average observed time series of carbon dioxide and temperature. So, carbon we have, so it is generally known that carbon dioxide emissions has a role behind global warming that is the general rise in the global temperature, but can we try to quantify it.

So, the detection and attribution methodologies have been developed over the years to delineate anthropogenic from natural drivers of climate change and impacts. Anthropogenic means the human induced. A majority of prior attribution studies which have used climate model simulations and observations or reanalysis datasets have found evidence for human induced climate change.

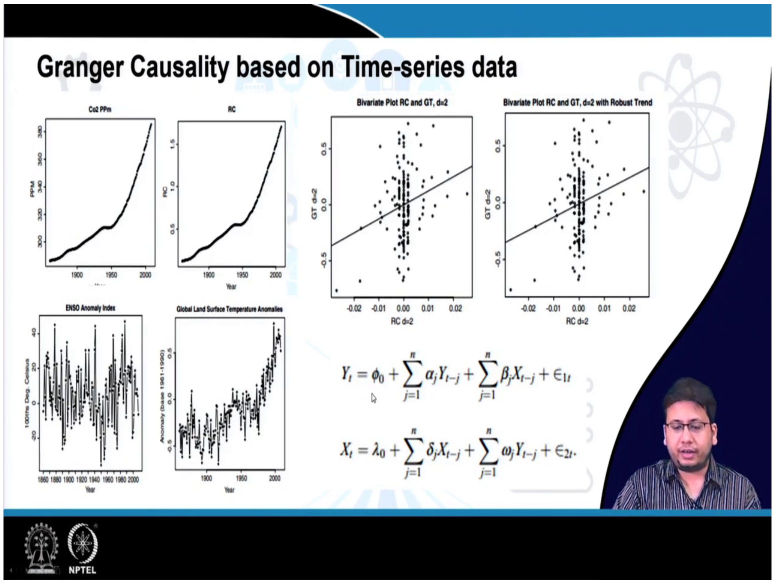
This paper tests the hypothesis that Granger causality can be extracted from the bivariate series of globally averaged land surface temperature observations and observed CO_2 in the atmosphere using a reverse cumulative Granger causality test. This the this proposed extension of the classic Granger causality test is better suited to handle the multisource nature of the data and provides for the statistical rigor. The results from this modified test show evidence for Granger causality from a proxy of total radiative forcings which in this case is our transformation of atmospheric CO_2 to global temperature.

Prior literature failed to extract these results via the standard Granger causality test. A forecasting test shows that the holdout set of the global temperature can be better predicted with the addition of lagged radiative forcing as a predictor lending further credibility to the Granger test results. However, since second order differenced RC is neither normally distributed nor variance stationary, caution should be exercised in the interpretation of our results.

So, basically the whole idea is that they have two main variables, the global temperature and CO_2 and they are like and they are relating it to another variable called the total radiative forcing which they denote by RC here. So, like so whenever we try to fit a Granger causal relations that is we try to express one variable in terms of the other, using a linear relation, we are basically making a hypothesis we.

And that is how hypothesis which we need to test using the data which we have had. So, it is basically a testing of hypothesis problem where we need to define a test statistic and then either and we have a null hypothesis and we either accept or reject the null hypothesis at certain level of confidence.

(Refer Slide Time: 20:16)



So, in this paper basically, they like the null hypothesis is that like one of the like. So, first of all these are some of the predictor variables. So, this is the thing which they are trying to predict, the global land surface temperature anomaly. So, as you can see it has like, it is steadily rising up like this which indicates the global warming. So, this is the variable which they are trying to predict or it is the effect variable for which they are looking for causes.

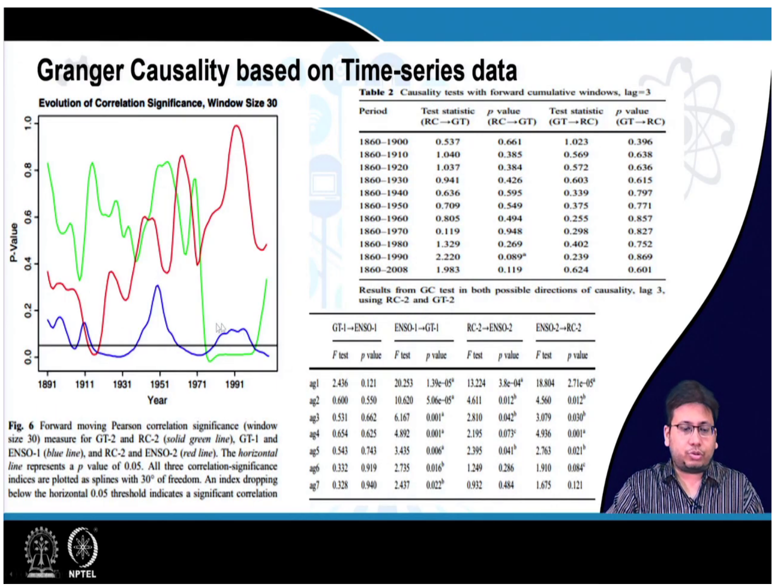
So, one possible cause is the CO_2 emission which also is increasing like this. And like if they find that it is some kind of a proxy for radiative forcing which also as you can see like its plot looks pretty much similar, although the units are different. And then there is something known as the ENSO anomaly index. So, like this is like a phenomena in the Pacific Ocean, which like which is like which can be understood as some kind of a proxy for warming. And so, the like they come up with these plots.

As you can see $d = 2$ means like a time lag of 2 time units. So, here they are plotting the this bivariate plot of radiative forcing and the global temperature. And this is what the plot looks like. So, like the when we go for Granger causality we will try to fix some kind of a linear relation between the two. But as you can see that it does not seem to be a that great an idea to represent this kind of data with a like a with a linear relation.

Similarly, this is like an like here they make another bivariate plot with what they known what is known as a robust trend. So, this is the hypothesis of Granger causality where like there are two variables Y and X. So, like let us say that X is the global temperature the thing which the so called effect and Y is the cause like which may be the CO_2 emissions and so on.

So, the grange causality relation is to express one like the present value of each one of them in term as a linear combination of the past values of both of the others. And then see, like if the these components the components of the other one in this case the ß's and the ω's if they should be zero or nonzero, ok.

(Refer Slide Time: 22:48)



And so like these are like, so like here for the like to understand what should be the proper lag, because as we had discussed earlier these effects cannot be instantaneous. It is not that if there is an emission this year, immediately the same year the temperature will also grow up go up, in fact, we like we can expect some kind of a lag relation. But how much should that lag be?

So, to they, so they like use the help or that what is known as the Pearson correlation coefficient like. So, and they and they actually measure the Pearson correlation coefficient between different pairs of variables, like the global temperature and RC, RC meaning the radiative forcing. And these two, this means the is something like the second derivative like they have like.

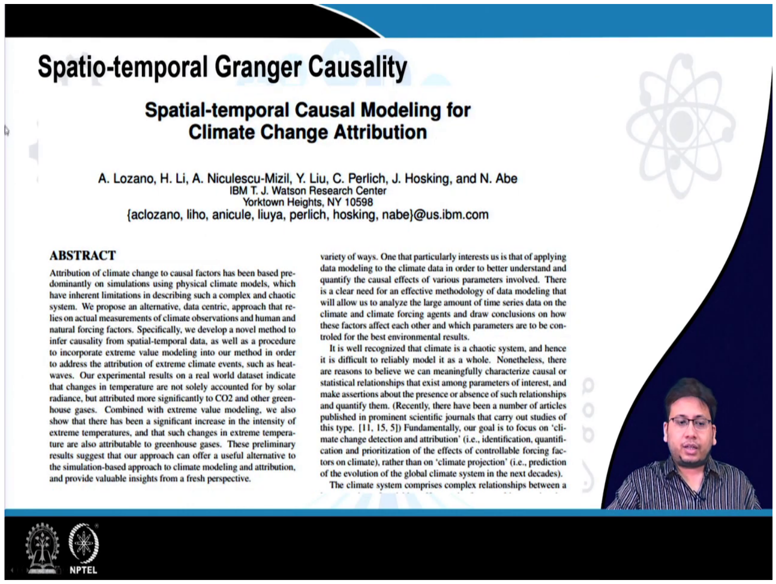
So, we have the original time series, then we can like take the time that the time derivative of it by doing first order differentiation. That is we, like we have, let us say we have a time series of X , we will a like X instead of $X(t)$ we will have $X(t - 1)$ and instead of $X(t - 1)$ we will have $X(t - 1) - X(t - 2)$ and so on. So, and this, so that is the first order differencing. Then, similarly we can have the second order differencing also.

So, like we have, so these things are done basically to get rid of the local oscillations and so on or to get rid of the local what is known as the noisy components and to make the series a bit smooth. So, after smoothing the series by this by tool second degree differentiating, they then they carry out the Pearson correlation test of the different time windows between the these, between the different pairs of variables.

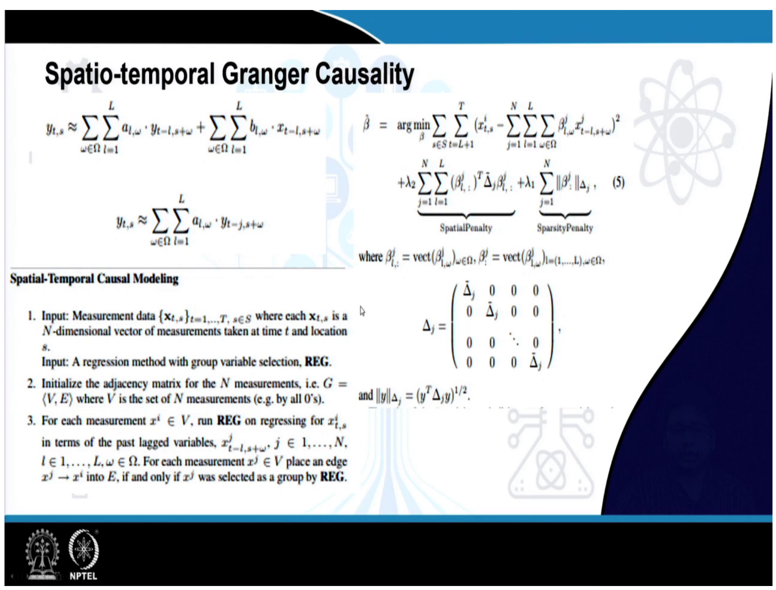
And they see that say around 1991, so the p value between GT-2 and RC-2 is like very high, it is like almost close to 1. While there are other periods of time when they like their correlation coefficient was it was effectively nil. So, like this shows the how the Pearson correlation coefficient between the two things or between the different pairs of things they vary. And as we can see that the for the same pair of quantities there might be periods when they are strongly correlated and there might be periods when they are not strongly correlated.

So, accordingly they try to come up with this kind of like time varying testing of the hypothesis. At different points of time or at different windows they like and it carry out the testing of the hypothesis using a suitable p value and see like at which time or which intervals of time the p value is high enough. That is the causal relation can be validated and in what which periods it is it does not hold. So, accordingly they come up with their conclusions.

(Refer Slide Time: 26:00)

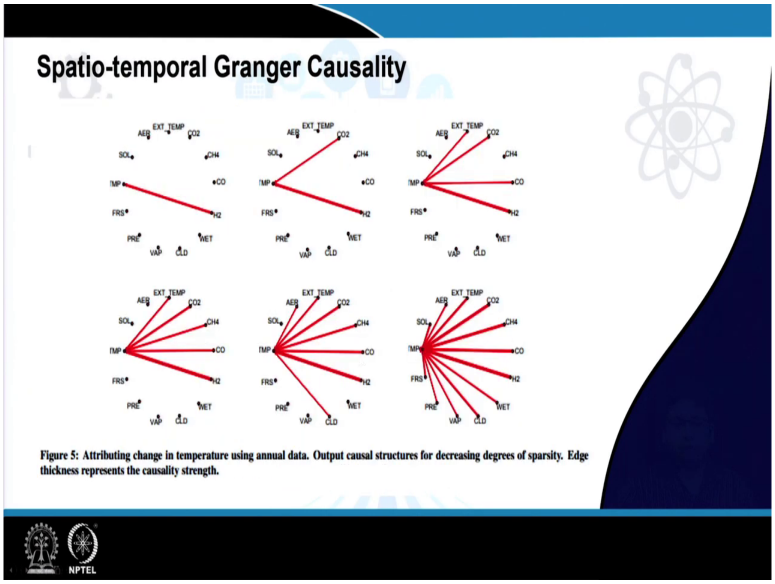


(Refer Slide Time: 26:20)



And then there here is another paper, this is also based on Granger causality, but here the aim is slightly different. Here the aim is to carry out spatiotemporal causal modeling for climate change attribution. Here unlike the previous works where we talked about the network, here again we have a graph, but the difference here is that the nodes here stand for the different variables, not locations, but variables.

(Refer Slide Time: 26:32)



And we try to come up with a network structure like there is a particular variable of interest which we like which we call as this TEMP, the temperature and so the aim here is to, so once again the aim is to attribute the warming that is the rise in global temperature. And we are trying to see which of the different factors like impact this global temperature more.

So, the so once again they take help of Granger causality with a catch. In this case, the target variable the temperature that is like again spatio-temporal. Let us say we are measuring it at a location s and a time t. So, when we are looking for predictors we will have the following condition on the predictors that we will look only in the vicinity.

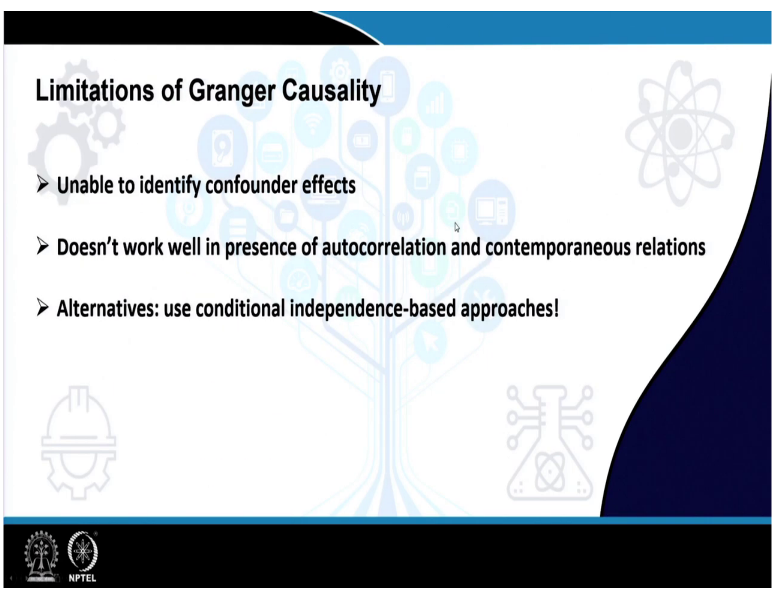
That is, for if the for like in case of the target variable if the location is s and time is t, then its predictors we will search like in a small neighborhood of the location s and within a short time window t like $t - 1$. And so like, so for that purpose they define some kind of a special penalty and a special penalty function, which means that if the if any predictor is far away like is located at a position which is far away from the target variable like it will be some kind of penalized. So, like I mean it can be accepted as a predictor, but with a penalty term.

So, the idea is that I do not want predictors which are very far apart. I want predictors which are close by. So, that kind of thing is ensured with the help of this special penalty function. Apart from that we also have a sparsity penalty function. This is the usual sparsity function which or the sparsity condition which we have in case of sparse regression where we do not want too many predictors, we only want the most critical predictors that is the variable selection problem.

And so accordingly they have these kind of a linear relation which they carry like, the linear regression we can call it as a regularized linear regression which they carry out and they get the final answer. And these are the output of the regression. So, here as you can see like when the it is very sparse, that is when the sparsity is very high then only there is only one edge in the network which shows that the H_2 the hydrogen concentration seems to be the strongest driver of the temperature.

Then, if we relax it a little bit then CO_2 becomes the next largest driver. Then, as we reduce the sparsity a little bit then all these things including the carbon monoxide etcetera these become the predictors and so on. And gradually, as we see as we keep on relaxing the sparsity we see more and more drivers being added. But the catch is that, like when the sparsity is very low then the most we see get only the most crucial drivers and those seem to be the hydrogen, carbon dioxide etcetera.

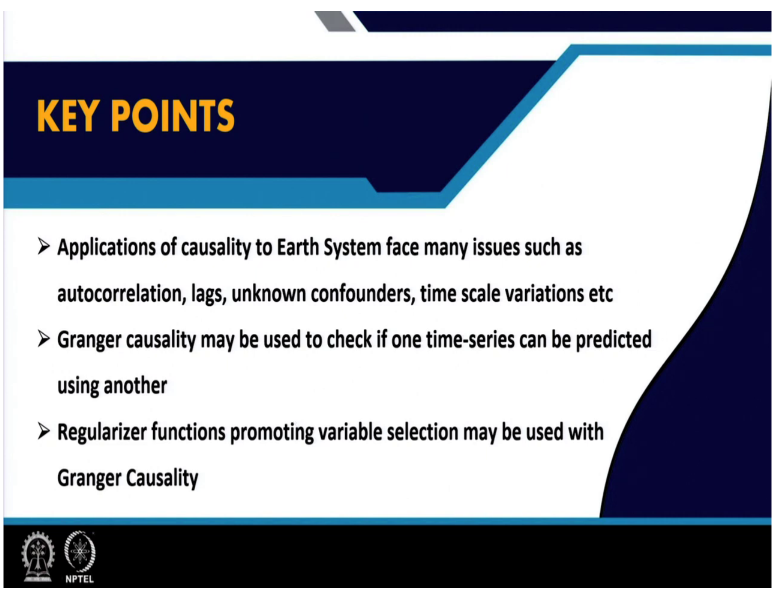
(Refer Slide Time: 29:48)



So, the basically, the limited the Granger causality also has some limitations namely that it is unable to identify the confounder effects. It does not work well in the presence of autocorrelation and contemporaneous relations. And so, the alternative to Granger causality is to use conditional independent based approaches which we will see in the next lecture.

(Refer Slide Time: 30:09)

(Refer Slide Time: 30:13)



So, these are the reference for the two papers that we discussed. And so, Applications of causality in Earth Science System Science may face many issues such as autocorrelations, lags, unknown confounders, time scale variations etcetera. Granger causality can be used to check if one-time-series can be predicted using another. And regularization functions can be used to promote variable, like variable selection under different criteria.

They can be used for Granger causality. So, that brings us to the end of this lecture where we will discuss more on the causality in the following lecture. Till then goodbye.