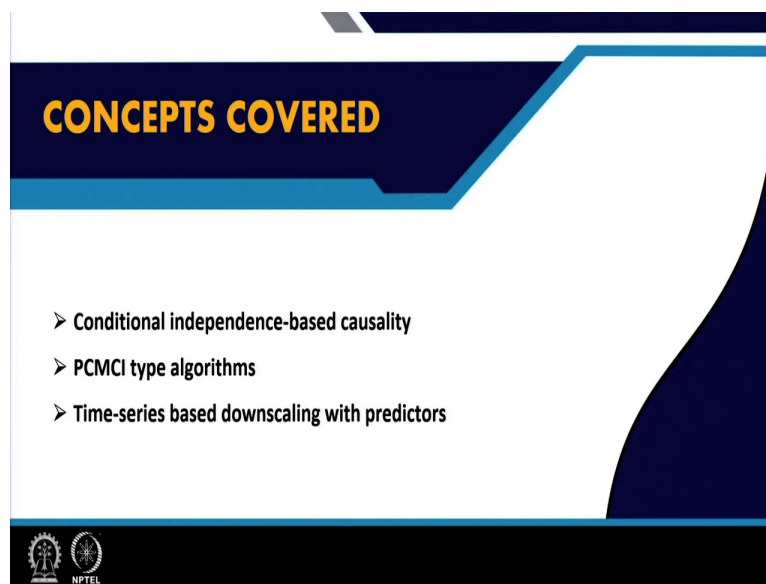


**Machine Learning for Earth System Sciences**  
**Prof. Adway Mitra**  
**Department of Computer Science and Engineering**  
**Centre of Excellence in Artificial Intelligence**  
**Indian Institute of Technology, Kharagpur**

**Module - 03**  
**Machine Learning for Discovering New Insights**  
**Lecture - 20**  
**Identifying Casual Relations from Time-Series - 2**

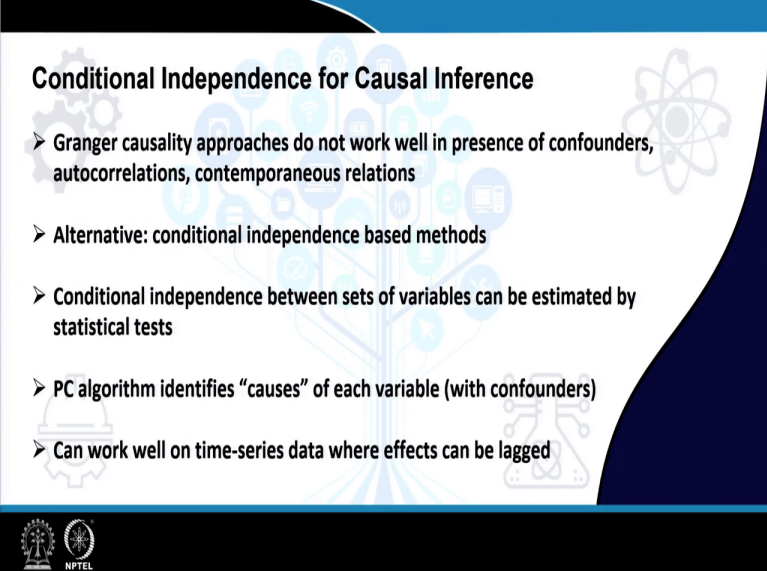
Hello everyone. Welcome to lecture 20 of this course on Machine Learning for Earth System Science. We are still in module 3, which is where we explore how Machine Learning can be used to Discover New Insights in Earth System Sciences. In the last lecture, we had been discussing about Identifying Causal Relations from Time Series. So, today also we will continue the same topic.

(Refer Slide Time: 00:48)



So, the concepts which we are going to cover today are conditional independence based causality, *PCMCI* types of algorithms, and time series based downscaling with predictors.

(Refer Slide Time: 01:00)



**Conditional Independence for Causal Inference**

- Granger causality approaches do not work well in presence of confounders, autocorrelations, contemporaneous relations
- Alternative: conditional independence based methods
- Conditional independence between sets of variables can be estimated by statistical tests
- PC algorithm identifies “causes” of each variable (with confounders)
- Can work well on time-series data where effects can be lagged

The slide features a background graphic of a tree with nodes, where each node contains a small icon representing different concepts like a gear, a person, a document, and a network. The slide is framed by a blue header and a dark blue footer containing the NPTEL logo.

So, like in the last lecture, we were discussing about Granger causality and its use cases, say to identify the causes of rising global temperature and so on, where we try to express the global time series of global temperature as a linear function of the time series of some other variables including the CO<sub>2</sub> concentration and so on.

Now, the Granger causality has certain drawbacks. It does not work well in the presence of confounders, autocorrelations in the time series and contemporaneous relations. Confounders mean like if they suppose there are multiple variables which we are focusing on, and like that is you will in case of Granger variable in Granger causality we are trying to express one variable as a linear function of the past values of the other.

But if there is another variable, there is a third variable which influences both the effect and the potential cause. In that case, Granger causality may not make much sense.

Then, like an alternative is to go for conditional independence based method which actually aim to identify these confounders and remove them with the help of the concept of conditional independence.

So, conditional independence between sets of variables can be estimated using statistical tests. So, like conditional independence is different from normal independence between two variables

in the sense that, like let us say we have three variables  $x$ ,  $y$  and  $z$ . So, we know that we can call  $x$  and  $y$  as independent of each other if like the joint distribution of  $x$  and  $y$  can be factorized as the marginal distributions of  $x$  and  $y$ , that is  $p(x, y) = p(x)p(y)$ .

But now if you consider the variable  $z$ , so like we can say that  $x$  and  $y$  are conditionally independent based on  $z$ , if p like if the joint, if the conditional joint distribution of  $x$  and  $y$  conditioned on  $z$ . It factorizes as the product of the conditional distribution of  $x$  on  $z$  and the conditional distribution of  $y$  on  $z$ , that is,  $p(x, y|z) = p(x|z)p(y|z)$ .

Now, independence of  $x$  and  $y$  and conditional independence of  $x$  and  $y$  based on  $z$  are two different things. One and one does not imply the other. So, like if  $x$  and  $y$  are conditionally independent based on  $z$ , then it is not impossible that  $z$  may play some kind of a like a confounder between as far as the relation between  $x$  and  $y$  are concerned.

Now, it might happen that is if we ignore  $z$  and if we look for the relation between  $x$  and  $y$ , we may find some kind of a relation by Granger causality, and we may think that  $x$  is the cause and  $y$  is the effect. But in reality it might happen that actually both of them are conditionally independent based on  $z$ , that is its  $z$  is a cause of  $x$  and  $z$  is also a cause of  $y$ .

So, these kind of if such a situation arises then Granger causality is not able to detect them in any way. But if we have these the conditional independence based approaches, then we can like estimate this or we can identify this situation and like then identify  $z$  as a potential confounder.

So, this the accordingly we have the *PC* algorithm which is known as like, it is named according to their inventors. So, *PC* algorithm identifies the cause of each variable with the confounders and it can work well on time series data where the effects are effects can be a lagged as well as they are contemporaneous. So, like Granger causality does not work at all on in a to a is unable to identify contemporaneous relations at all, that is for the following reason that when I am like let us say that I am trying to express  $x(t)$  as a function of  $y(t)$ .

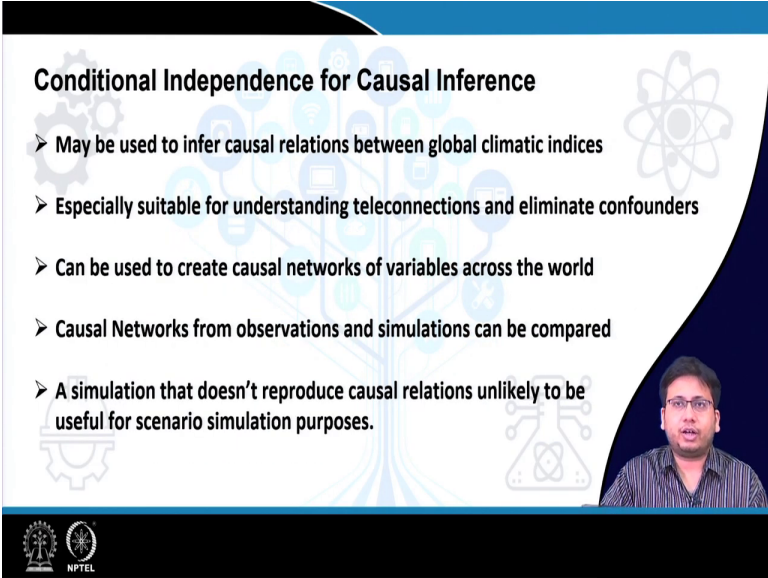
Now, suppose I have able to do it, suppose I turn does turn out that  $x(t)$  is equal to like some a times  $y(t)$  plus other things. Now, even if that is the case, then we will never be sure whether to say that  $x(t)$  is the cause or  $y(t)$  is the cause;  $x(t)$  is the effect or  $y(t)$  is the effect. That is

because if  $x(t)$  can be expressed as a linear function of  $y(t)$ , then  $y(t)$  can also be expressed as a linear function of  $x(t)$ . So, we cannot understand any cause effect relation in that case.

Apart from contemporaneous relations like if  $x(t)$  could be expressed as in terms of past values of  $y(t)$ , then we could say that  $y$  is cause and  $x$  is effect because the cause must precede the effect. But if in case of contemporaneous relations, we that like there is no such constraint of related to time. So, we are not able to even if we can find a contemporaneous relation, it will does not help us to understand which is cause and which is effect.

Now, this drawback of Granger causality is solved at least partially by the *PC* family of algorithms.

(Refer Slide Time: 06:36)



**Conditional Independence for Causal Inference**

- May be used to infer causal relations between global climatic indices
- Especially suitable for understanding teleconnections and eliminate confounders
- Can be used to create causal networks of variables across the world
- Causal Networks from observations and simulations can be compared
- A simulation that doesn't reproduce causal relations unlikely to be useful for scenario simulation purposes.

The slide features a background with faint icons of a gear, a lightbulb, and a network diagram. A small video inset in the bottom right corner shows a man with glasses speaking. The NPTEL logo is visible in the bottom left corner.

So, here this kind of conditional independence of based causal inference, this can be used to infer causal relations between global climate indices, so in different parts of the world. Say, for example, the ENSO index which is related to the Pacific Oceans temperature or the ISMR the Indian monsoon that is another index. And there are various other indices in the world like related to climate various climatic variables in different parts of the world.

Especially, this is suitable for understanding that teleconnections and to eliminate the confounders. And this kind of approach can also be used to create causal networks of variables

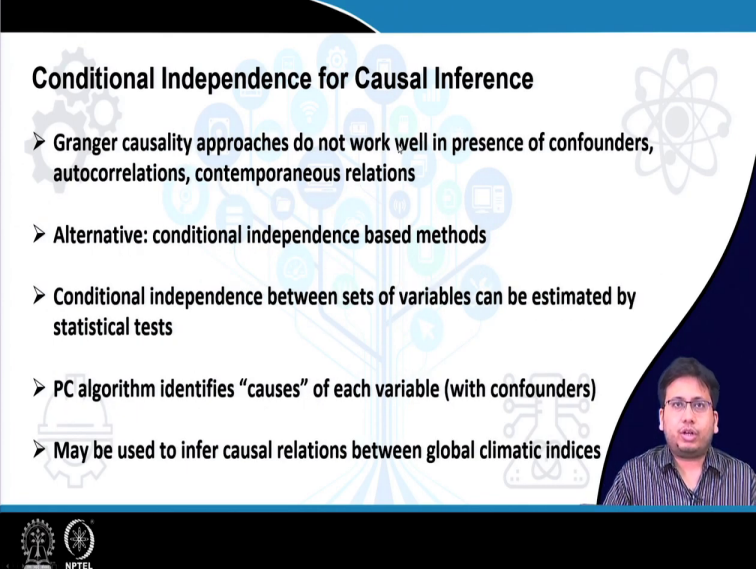
across the world. And these causal networks can also be constructed not only based on observations, but also on simulations. And then it is possible to compare the different causal networks also.

The need for such comparison is that a simulation that does not reproduce the causal relations is unlikely to be very useful for the purpose of simulating different future scenarios.

So, it is necessary to see if whenever there is a model which simulates the global climate or something like that, it is necessary to see whether it is able to like, is it just simulating some kind of like variables or does it actually have the causal relations that are observed in the real world. Because if that is not there then there is no point in using that model for future simulations.

The future simulation we want to do like, that is if we are interested in asking questions like if the global temperature rises so much then what will be the effects of it or if the carbon emission goes up like this then what are going to be the effects of this. So, this can be, model can be able to answer this question meaningfully only if it is able to capture the causal relations well.

(Refer Slide Time: 08:55)



**Conditional Independence for Causal Inference**

- Granger causality approaches do not work well in presence of confounders, autocorrelations, contemporaneous relations
- Alternative: conditional independence based methods
- Conditional independence between sets of variables can be estimated by statistical tests
- PC algorithm identifies “causes” of each variable (with confounders)
- May be used to infer causal relations between global climatic indices

The slide features a background with faint icons of a gear, a lightbulb, and a network diagram. A small video inset in the bottom right corner shows a man with glasses speaking. The NPTEL logo is visible in the bottom left corner.

So, it is very necessary to like understand the causal structure or the causal relationships which are any particular simulation model entails. So, one way to do it is through this kind of causal networks.

(Refer Slide Time: 09:15)

## Causal Inference based on Conditional Independence

**Causal Discovery for Climate Research Using Graphical Models**

IMME EBERT-UPHOFF  
*Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado*

YI DING  
*School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia*

(Manuscript received 12 July 2011, in final form 4 February 2012)

ABSTRACT

Causal discovery seeks to recover cause-effect relationships from statistical data using graphical models. One goal of this paper is to provide an accessible introduction to causal discovery methods for climate scientists, with a focus on constraint-based structure learning. Second, in a detailed case study constraint-based structure learning is applied to derive hypotheses of causal relationships between four prominent modes of atmospheric low-frequency variability in boreal winter including the Western Pacific Oscillation (WPO), Eastern Pacific Oscillation (EPO), Pacific-North America (PNA) pattern, and North Atlantic Oscillation (NAO). The results are shown in the form of static and temporal independence graphs also known as Bayesian Networks. It is found that WPO and EPO are nearly indistinguishable from the causal perspective as strong simultaneous coupling is identified between the two. In addition, changes in the state of EPO (NAO) may cause changes in the state of NAO (PNA) approximately 18 (3-6) days later. These results are not only consistent with previous findings on dynamical processes connecting different low-frequency modes (e.g., interaction between synoptic and low-frequency modes) but also provide the basis for formulating new hypotheses regarding the time scale and temporal sequencing of dynamical processes responsible for these connections. Last, the authors propose to use structure learning for climate networks, which are currently based primarily on correlation analysis. While correlation-based climate networks focus on similarity between nodes, independence graphs would provide an alternative viewpoint by focusing on information flow in the network.

**Algorithm 2.1: The PC-simple algorithm [5]**

Input:  $D$ , a data set for the set of predictor variables  $X = \{X_1, X_2, \dots, X_n\}$  and the target variable  $Z$ ; and  $\alpha$ , significance level for conditional independence tests.  
Output:  $PC$ , the subset of  $\{X_1, X_2, \dots, X_n\}$  that comprises parents and children of  $Z$

```
1: let  $k = 0$ 
2: let  $PC^k = \{X_1, X_2, \dots, X_n\}$ 
3: while  $|PC^k| > k$  do
4:   let  $k = k + 1$ 
5:   let  $PC^k = PC^{k-1}$ 
6:   for each  $X \in PC^{k-1}$  do
7:     for each  $S \in PC^{k-1} \setminus \{X\}$  and  $|S| = k - 1$  do
8:       if  $X$  and  $Z$  are independent given  $S$  at significance level  $\alpha$ 
9:         let  $PC^k = PC^k \setminus \{X\}$ 
10:  end for
11: end for
12: end while
13: output  $PC^k$ 
```

NPTEL

Now, Granger causality as we had said earlier is unable to deal with this kind of phenomena. So, that is why we use this conditional based, conditional independence based approaches for this purpose of constructing these global climate networks or sorry I mean global causal networks and comparing them.

So, here is like this paper appeared in about in 2012, quite some time back about causal inference based on conditional independence. So, here like they consider four major indices as I was just talking about this. So, the causal discovery seeks to recover cause effect relationships from statistical data using graphical models.

So, graph, remember graphical models we have already talked about. It is a graph where the different variables each variable represents a random variable. One goal of this paper is to provide an accessible introduction to causal discovery methods for climate scientist with a focus on constraint based structure learning. By structure they mean the structure of the causal network.

Second in a detailed case study constraint, constraint based structure learning is applied to derive hypothesis of causal relationships between 4 prominent modes of atmospheric low frequency variability in boreal winter. So, these are the climate indices they are talking about. First of all

the WPO or West Pacific Oscillation, then East Pacific Oscillation or EPO, then Pacific North America PNA pattern and the North Atlantic Oscillation NAO.

Now, the relation between these variables these are like roughly known to atmospheric scientists who have some intuition about the physics of the process and they know which variable influences, which other variable. The results are shown in the form of static and temporal independence graphs also known as Bayesian networks.

It is found that WPO and EPO are nearly indistinguishable from the cause effect perspective as strong simultaneous coupling is identified between the two. In addition, changes in the state of either EPO or NPO may cause changes in the state of NAO or PNA correspondingly approximately, like 18 days later. In case of NAO and PNA, this lag period it becomes 3 to 6 days.

These results are not only consistent with previous findings on dynamical processes connecting different low frequency modes, but also provide the basis for formulating new hypothesis regarding the time scale and temporal sequencing of dynamical processes responsible for these connections.

Last, the authors propose to use structural learning for climate networks which are currently based primarily on correlation analysis. While correlation based climate networks focus on similarity between nodes, independence graphs would provide an alternative viewpoint by focusing on information flow in the network.

So, this is the PC like the PC algorithm which we talked about. This PC algorithm they have used for the structure learning which is mentioned repeatedly in the abstract. So, the PC algorithm, so this is the simplest or the most basic version of the PC algorithm. There are more sophisticated versions of it as we will see later in this lecture.

So, the input here is a set of predictor variables  $X$ , let us say there are small  $m$  number of such variables and in for each of them we have a time series of observations. And we have a target variable  $Z$ . That is the like  $Z$  is the my observation which we like for which I am trying to look for potential causes.

And we also have a significance level for the causal independence test which is called  $\alpha$ . Like, so like we know that for this convey this test or hypothesis testing, we need to have some level of significance based on which we can accept or reject the null hypothesis. So, that is the  $\alpha$ .

So, this is how the algorithm works. So, the basic idea is that like they consider different sets of variables and like a like a let that set  $S$ , that is of different sizes. And they take every single predictor  $X$  and they see whether  $X$  is independent of the target variable  $Z$  or not given that particular  $S$ . And that for when they are testing for this independence they use various statistical tests for conditional independence between different variables, and that those tests, they required the significance level  $\alpha$ .

So, and if they, like if  $X$  and  $Z$  are found to be independent given  $Z$ , then  $X$  is removed from the potential predictors of  $Z$  and so on. So, like finally, what we are left; so, like step by step all these like, each of the variables get removed or are retained based on the results of this kind of conditional independence test. And this conditional independence test must be done for like different sets of variables of different sizes.

And so, like ultimately what we are left with is a set or a subset of these initial variables, the potential causes and those that small subset which is left over after this algorithm, these are supposed to be the causes of the variable of the target variable  $Z$ .

And again these kind of things, like if we are doing the lagged relation then each of these variables  $X$ , they  $X_1, X_2, \dots, X_m$  they must be considered at different lag values. So, in the study which they have done, this is the graphical model they have come up with as the output. So, like this each of these nodes and for the different indices they mentioned.

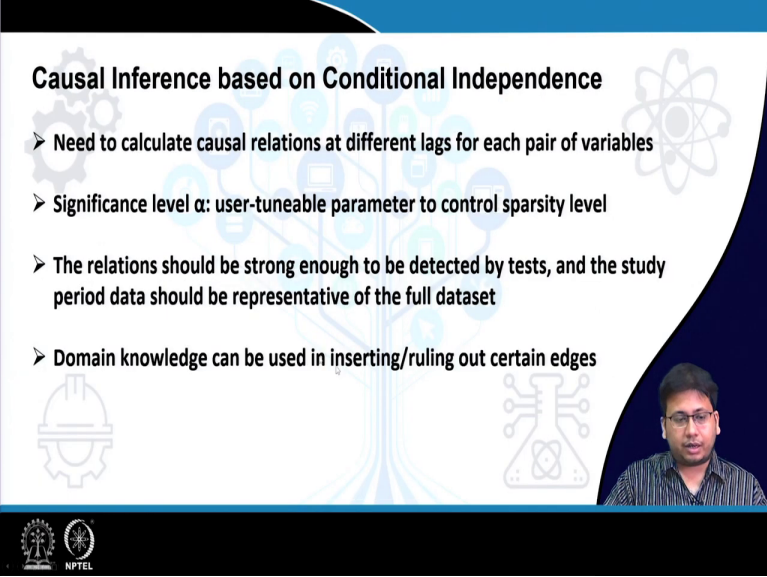
So, these arrows, these indicate like the cause effect relation. Like for example, this arrow it indicates the like a that is this is the cause and this is the effect. And each of these arrows they are associated with some kind of a time period which means that the lag.

So, like here it seems that EPO will be influencing NAO at like at with a lag of about 18 days. Similarly, WPO has a week weekly influences this PNA at a lag of 9 days. Similarly, there are



these 0 they indicate contemporaneous relations that is the cause effect relations take place simultaneously.

(Refer Slide Time: 16:33)



**Causal Inference based on Conditional Independence**

- Need to calculate causal relations at different lags for each pair of variables
- Significance level  $\alpha$ : user-tuneable parameter to control sparsity level
- The relations should be strong enough to be detected by tests, and the study period data should be representative of the full dataset
- Domain knowledge can be used in inserting/ruling out certain edges

The slide features a blue header and footer. The footer contains the NPTEL logo and a small video inset of a man with glasses speaking. The background has faint icons of a gear, a lightbulb, and a network diagram.

And now causal independence based on conditional independence in this case we need to calculate the causal relations at different lags for each pair of variables. Then, there is the significance level  $\alpha$ , this is a user-tunable parameters. And so, the relations should be strong enough to be detected by the tests, and the study a study period should be representative of the full data set.

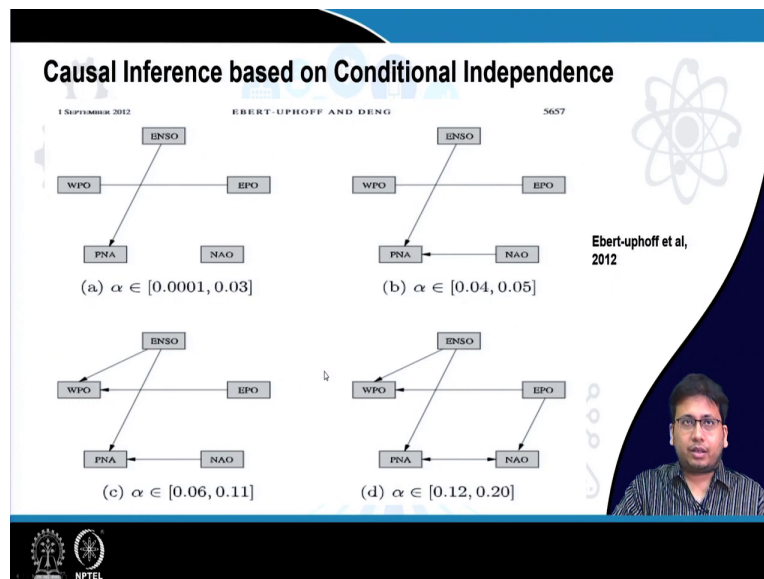
So, now like if their relation is weak that is it may in that case depending on our  $\alpha$  it may not be detected by the test, that is the hypothesis may get rejected. So, like that is again this is somewhat like the sparsity parameter which we talked about in the Granger causality. Just like for high values of sparsity, we have only a small number of variables they pass the causality test, and hence we have a very sparse graph.

So, in similarly in this case also depending on the this significance level  $\alpha$ , either a very few or many of the potential predictors may pass this condition, may pass or fail this conditional independence test. So, this significance level  $\alpha$  this largely determines how strong or how sparse this graph is going to be.

And in some situations we may have some kind of domain knowledge. That is we may already know by the according to the our knowledge of physics which climate scientists or other domain scientists may have, they can provide us some kind of clues that these two variables definitely have a relation and this is a cause and this is an effect or it might be negative also that these two variables. They do not have a relation, so which they may have found out by other methods.

So, if there are such domain knowledge is available, they it can be like incorporated into the graphs either by forcefully inserting or by ruling out certain edges.

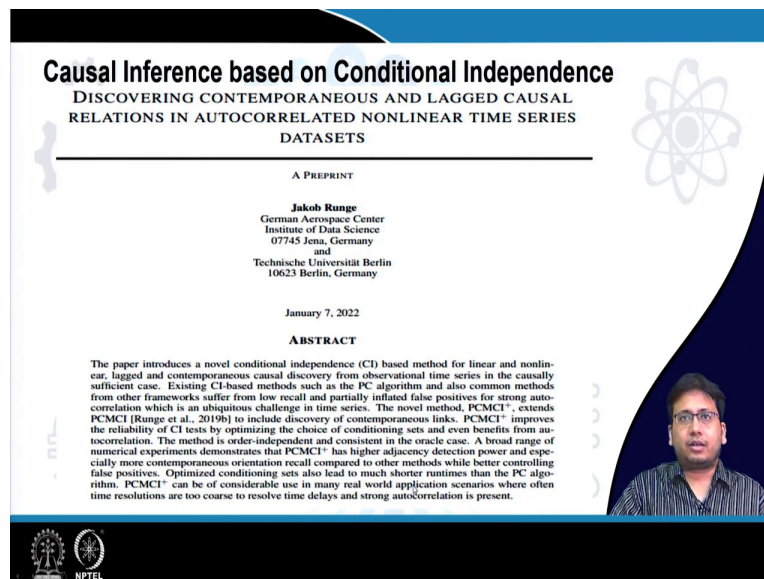
(Refer Slide Time: 18:40)



So, like, so this  $\alpha$  as I said this largely depend or influences how sparse the graph is going to be. So, in this case, in the paper which we are discussing if for very small values of  $\alpha$  like this, then we have only two edges in the graph. So, this edge as you can see it is like its an undirected edge which means that we see that there is some kind of a relation between them, but we are unable to identify the I mean which is the cause and which is the effect.

And then, for slightly higher values of  $\alpha$  we may see that like a few more edges may pass the test. And then, as we can as we go on increasing the value of  $\alpha$  we see that more and more edges are being inserted as the conditional independence because that is impacting the results of the conditional independence tests.

(Refer Slide Time: 19:40)



So, now, this same concept this can be extended for more for other purposes also and also for making them more robust to certain specific situations such as the contemporaneous and lagged relations. So, this paper by Jakob Runge who is one of the foremost scientists working on the domain of causality right now. So, this paper, so what it does is it proposes an advanced version of the PC algorithm which we mentioned earlier.

The paper introduces a novel conditional independence based method for linear and non-linear lagged and contemporaneous causal discovery from observational time series in the causally sufficient case. Existing CI-based methods such as the PC algorithm and also common methods from other frameworks like Granger causality suffer from low recall and partially inflated false positives for strong autocorrelation which is an ubiquitous challenge in time series, right.

So, like I have we have already discussed in the previous lecture about the presence of autocorrelations, like in case of in case auto autocorrelations are strong. Like we may end up making some kind of mistakes in our estimation of the causal relations.

The novel method which they call as  $PCMCI^+$  extends the  $PCMCI$  to include discovery of contemporaneous links.  $PCMCI^+$  improves the reliability of the conditional independence tests by optimizing the choice of conditioning sets and even benefits from autocorrelation.

So, been autocorrelation earlier was seen found to be detrimental to identifying causal relations and they were they often resulted in mistakes, but here the claim is that this autocorrelation will actually be utilized to make a more efficient algorithm.

The test the method of is order independent and consistent in the oracle case. A broad range of numerical experiments demonstrates that  $PCMCI^+$  has higher adjacency detection power and especially more contemporaneous orientation recall compared to other methods while better controlling false positives. So, false positive as you know means that there is no relation, but there is no causally relation, but your algorithm mistakenly identifies that there is a causal relation.

And optimized conditioning sets also lead to much shorter runtimes than the PC algorithm.  $PCMCI^+$  can be of considerable use in many real world application scenarios where often time resolutions are too coarse to resolve time delays and strong autocorrelation is present.

Like, so like if you see here in the PC algorithm here like they are taking the set  $S$  of the or the like I mean the conditioning set and for different set of these sets they are carrying out the conditional independence test. So, this step of course, like is like this set  $S$  they it can be a like any set of any size. So, like there is obviously, a large space it covers. So, they are saying that they in  $PCMCI$  algorithm, they will navigate this smartly and so that only a smaller number of this as variable needs to be chosen.

(Refer Slide Time: 23:20)


### Causal Inference based on Conditional Independence

---


**Algorithm 1** (PCMCI<sup>+</sup> / PCMCI<sub>0</sub><sup>+</sup> lagged skeleton phase)



**Require:** Time series dataset  $X = (X^1, \dots, X^N)$ , max. time lag  $\tau_{\max}$ , significance threshold  $\alpha_{PC}$ , CI test  $CI(X, Y, Z)$  returning  $p$ -value and test statistic value  $I$

- 1: **for all**  $X_t^j$  in  $X_t$  **do**
- 2:   Initialize  $\hat{B}_t^-(X_t^j) = X_t^- = (X_{t-1}^1, \dots, X_{t-\tau_{\max}}^1)$  and  $I^{\min}(X_{t-\tau}^i, X_t^j) = \infty \ \forall X_{t-\tau}^i \in \hat{B}_t^-(X_t^j)$
- 3:   Let  $p = 0$
- 4:   **while** any  $X_{t-\tau}^i \in \hat{B}_t^-(X_t^j)$  satisfies  $|\hat{B}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}| \geq p$  **do**
- 5:     **for all**  $X_{t-\tau}^i$  in  $\hat{B}_t^-(X_t^j)$  satisfying  $|\hat{B}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}| \geq p$  **do**
- 6:        $S =$  first  $p$  variables in  $\hat{B}_t^-(X_t^j) \setminus \{X_{t-\tau}^i\}$
- 7:        $(p\text{-value}, I) \leftarrow CI(X_{t-\tau}^i, X_t^j, S)$
- 8:        $I^{\min}(X_{t-\tau}^i, X_t^j) = \min(|I|, I^{\min}(X_{t-\tau}^i, X_t^j))$
- 9:       **if**  $p\text{-value} > \alpha_{PC}$  **then** mark  $X_{t-\tau}^i$  for removal
- 10:    Remove non-significant entries and sort  $\hat{B}_t^-(X_t^j)$  by  $I^{\min}(X_{t-\tau}^i, X_t^j)$  from largest to smallest
- 11:    Let  $p = p + 1$
- 12: **return**  $\hat{B}_t^-(X_t^j)$  for all  $X_t^j$  in  $X_t$



Runge, 2022



So, this is the  $PCMCI^+$  algorithm which is the sophisticated form of the  $PCMCI$  algorithm.

So, as you can see the times input here is again the time series data set  $X$  of the different predictor variables  $N$ . Then, there is the maximum time lag  $\tau_{\max}$  which is acceptable. Then, there is a just like we had the  $\alpha$  the significance threshold in the CI in the PC algorithm, in this case also we have the same  $\alpha^{PC}$ .

We have; apart from that we need a CI test, a conditional independence test involving like two variables  $X$  and  $Y$ , and a set and a conditioning set  $Z$  which returns a  $p$  value and test statistic value. So,  $p$  value like basically it means the probability of accepting the null hypothesis that  $X$  and  $Y$  are independent given  $Z$ .

So, like, so basically this is how the algorithm works. I will not go into the details of the algorithm, but so, but once again the basic idea is that it takes us of this set  $S$  the conditioning set it which is done in a smart way using the various using the data structure.

And like once that conditioning set is chosen then we do this CI analysis. So, note that note the presence of this  $t - \tau$  that is they are actually calculating the lagged independence relations. That is instead of comparing  $X(t)$  and  $Y(t)$ , they are considering  $X(t - \tau)$  and  $Y(t)$ , like to

take care of the lagged, I mean the I mean the lagging effects. And so, this like as a result they get the value of the test statistic and also the p value.

Now, if that p value is greater than the  $\alpha$ ; that means, that the condition they can accept the null hypothesis that they are there is that  $X$  and  $Y$  are indeed independence given up given for this particular conditioning set  $S$ .

So, that basically means that  $X$  is not a predictor of  $Y$ . So, it can be removed; for as I mean not  $X$ , I mean  $X(t - \tau)$ ;  $X$  at that particular lagged value that is not a predictor of  $Y$ . So, we can remove it from the list of potential predictors, and we can turn on. And every time as we keep on iterating we just keep on refining the this set, we keep on shrinking it, so that we have to do lesser searches over this conditioning set.

(Refer Slide Time: 26:08)

**Model evaluation based on Causal Relationships**

<https://doi.org/10.1038/s41467-020-15195-y> OPEN

### Causal networks for climate model evaluation and constrained projections

Peer Nowack<sup>1,2,3,4,5</sup>, Jakob Runge<sup>5,1</sup>, Veronika Eyring<sup>6,7</sup> & Joanna D. Haigh<sup>1,2</sup>

Global climate models are central tools for understanding past and future climate change. The assessment of model skill, in turn, can benefit from modern data science approaches. Here we apply causal discovery algorithms to sea level pressure data from a large set of climate model simulations and, as a proxy for observations, meteorological reanalyses. We demonstrate how the resulting causal networks (fingerprints) offer an objective pathway for process-oriented model evaluation. Models with fingerprints closer to observations better reproduce important precipitation patterns over highly populated areas such as the Indian subcontinent, Africa, East Asia, Europe and North America. We further identify expected model interdependencies due to shared development backgrounds. Finally, our network metrics provide stronger relationships for constraining precipitation projections under climate change as compared to traditional evaluation metrics for storm tracks or precipitation itself. Such emergent relationships highlight the potential of causal networks to constrain long-standing uncertainties in climate change projections.

The slide includes a stylized atomic symbol logo on the right and a small video inset of a speaker in the bottom right corner. Logos for the Indian Institute of Space Science and Technology (IIST) and NPTEL are visible at the bottom left.

So, now we already talked about this causal networks for evaluating the climate model simulations. So, causal networks for open for climate model evaluation and constraint projections. So, this is a like another paper which appeared very recently. So, global climate models are central tools for understanding past and future climate change.

The assessment of model skill in turn can benefit from modern data science approaches. Here we apply causal discovery algorithms to sea level pressure data from a large set of climate model simulations and as a proxy for observations meteorological reanalyses.

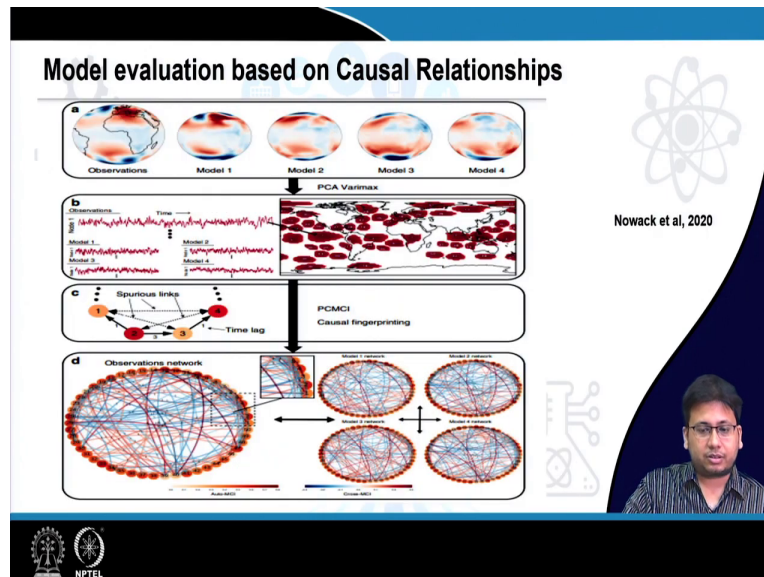
So, like reanalyses is basically the like a combination of observations and the and simulations by different models. So, so this is done with the help of data assimilation. We have already talked about data assimilation in one of our earlier lectures. So, when a climate or when any process model runs and it is like it is also calibrated with data at regular intervals of time. So, that the simulations it produces are the are actually in sync with the with the real world measurements, that is called as reanalysis data.

We demonstrate how the resulting causal networks or the fingerprints as they can call it, they offer an objective pathway for process oriented model evaluation. Models with fingerprints closer to observations better reproduce important precipitation patterns with over highly populated areas such as the Indian Subcontinent, Africa, East Asia, Europe and North America.

We further identify expected model inter dependencies due to shared development backgrounds. The; by that they mean that these models or the different process models which they are comparing many of them actually have certain components or certain sub models in common and as a result many of them may share certain biases and so on.

Finally, our network metrics provide stronger relationships for constraining precipitation object, precipitation projections under climate change as compared to traditional evaluation metrics for storm track or precipitation itself. Such emergent relationships highlight the potential of causal networks to constrain long standing uncertainties in climate change predictions.

(Refer Slide Time: 28:54)



So, like let us say these are the observations which we have of different earth system variables in different parts of the earth, and these are the simulated the earth systems simulated by 4 different models.

So, like for each of them we have measurements of the different variables. So, like which we can call as the time series. So, this is the like let us say this is the observational time series of a particular variable and these are the simulate the simulated time series.

So, so first of all they do some kind of like PCA algorithm to reduce the dimensionality and so on, and then they use this the *PCMCi* algorithms to build the network. So, like these patches which you are saying each of these are numbered, so each of these patches actually are some kind of a index. And or like we can call these or we can understand these as a variables, and so each for each of these variables we have a time series.

So, like, so based on this they apply the *PCMCi* algorithm, and I like construct the kind of causal network that they mentioned by like adding an a directed edge from one variable to another, provided they have some relation between, provided they are not independent of each other. And so they finally, come up with some kind of a fingerprint or a causal network like this, so where they have these kinds of time lags and so on.

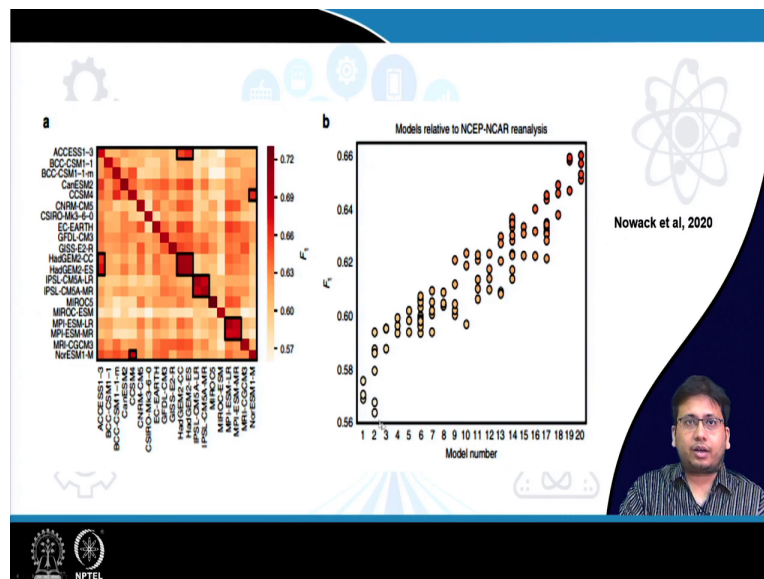


And once they; so, this is like the observation network which they might be getting. So, these nodes these are the different variables at certain periods of time. And then there are the like they we find that edges between different pairs of them and each of these edges may have some lags also. And similarly, for each of the simulations also we will find similar networks, and then we have to compare the different networks.

Now, there are different ways or different metrics of comparing two networks, like we can compare edge wise or we can like do component wise, like also we can compare the different properties of networks.

So, if you remember in our earlier lecture on networks, we had actually talked about some various like connectivity and various other things. So, all these structures can be the community structures of the different network, they can be identified and so on. So, this paper they actually discuss a host of these network based metrics in which they compare the observational network as well as the simulation network.

(Refer Slide Time: 31:43)



And accordingly here they for. So, these are the different climate models which they have compared. So, these GFDL, CM3, EC Earth etcetera these are various models which participate in the CMIP project, the coupled model inter inter-comparison project, where the models

developed by various agencies in the world, they are compared and then their relations with the actual observations here are plotted in this heat map. So, this that is with the help of their corresponding network structure.

So, they have this F1 score. So, F1 score like basically it is related to the precision and recall and they show here that the higher the higher for the different models, they calculate this F1 score and the higher the score, that score is it indicates a greater degree of similarity with the of the of that simulations network with the observational network.

(Refer Slide Time: 32:49)



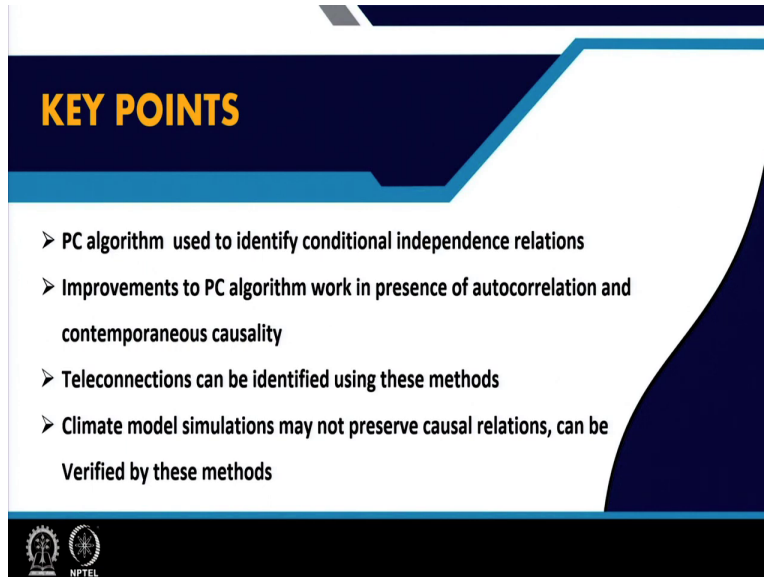
## REFERENCES

- Ebert-Uphoff I, Deng Y. Causal discovery for climate research using graphical models. *Journal of Climate*. 2012 Sep 1;25(17):5648-65.
- Runge J. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence 2020* Aug 27 (pp. 1388-1397). PMLR.
- Nowack P, Runge J, Eyring V, Haigh JD. Causal networks for climate model evaluation and constrained projections. *Nature communications*. 2020 Mar 16;11(1):1-1.



So, these are the references of the different three different papers which we discussed today.

(Refer Slide Time: 32:53)



**KEY POINTS**

- PC algorithm used to identify conditional independence relations
- Improvements to PC algorithm work in presence of autocorrelation and contemporaneous causality
- Teleconnections can be identified using these methods
- Climate model simulations may not preserve causal relations, can be Verified by these methods

NPTEL

The key points to be taken home from this lecture are that first of all the PC algorithm can be used to identify conditional independence relations. And the then improvements to the PC algorithm were like they work in presence of autocorrelation as well as the contemporaneous causality. And teleconnections can be identified using these methods. And the model simulations which may not preserve causal relations they also can be identified by these methods.

So, that brings us to the end of this lecture. In the following lectures, we will see some other applications of machine learning to find new knowledge in the Domain of Earth System Science.

So, till then goodbye.