Benchmarking of Data-Driven Causality Discovery Approaches in the Interaction between Arctic Sea Ice and Atmosphere

Presented by CyberTraining 2020 Team 6:

Yiyi Huang[1], Matthäus Kleindessner[2], Debvrat Varshney[4], Alexey Munishkin[3]

RA: Pei Guo[4]

Faculty: Jianwu Wang[4]

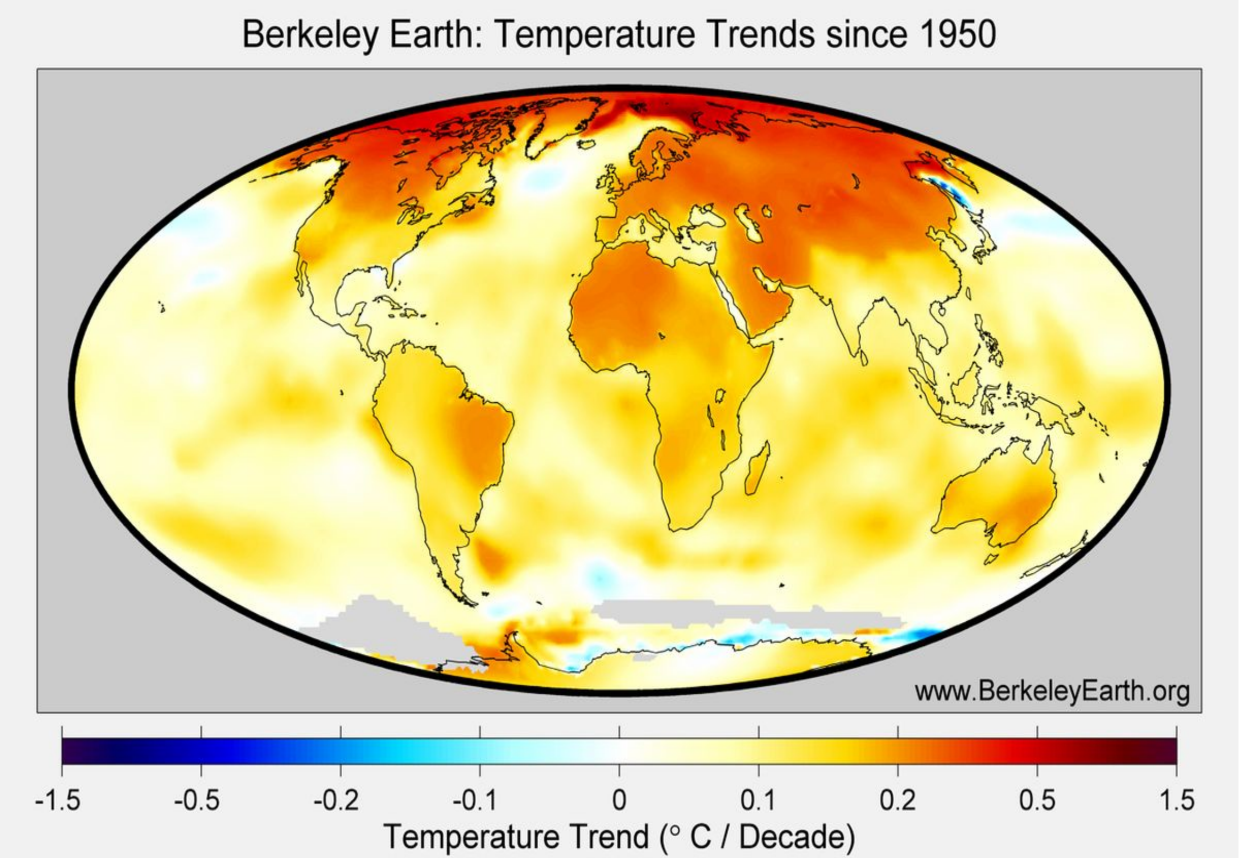
1. Department of Hydrology and Atmospheric Sciences, University of Arizona
2. School of Computer Science & Engineering, University of Washington
3. Department of Computer Science & Engineering, University of California, Santa Cruz
4. Department of Information Systems, UMBC

UMBC CyberTraining: http://cybertraining.umbc.edu/

Table of Contents

- Motivation: Discover relationships between the atmosphere and sea-ice
- Data: Thermodynamic and Dynamic (atmosphere variables) factors
  - Collected from as far back as 1978 and various centers
  - Different variables and data-sets
- Pre-processing of Data:
  - Time-series data that is decomposed and normalized
  - Additional steps for i.d.d. Causal discovery methods
- Causal discovery methods: TCDF, NOTEARS, DAG-GNN
- Results: causal discovery graphs and hyperparameter sensitivity analysis
- Conclusion and References:
  - A good first step and interesting results but more research is needed...

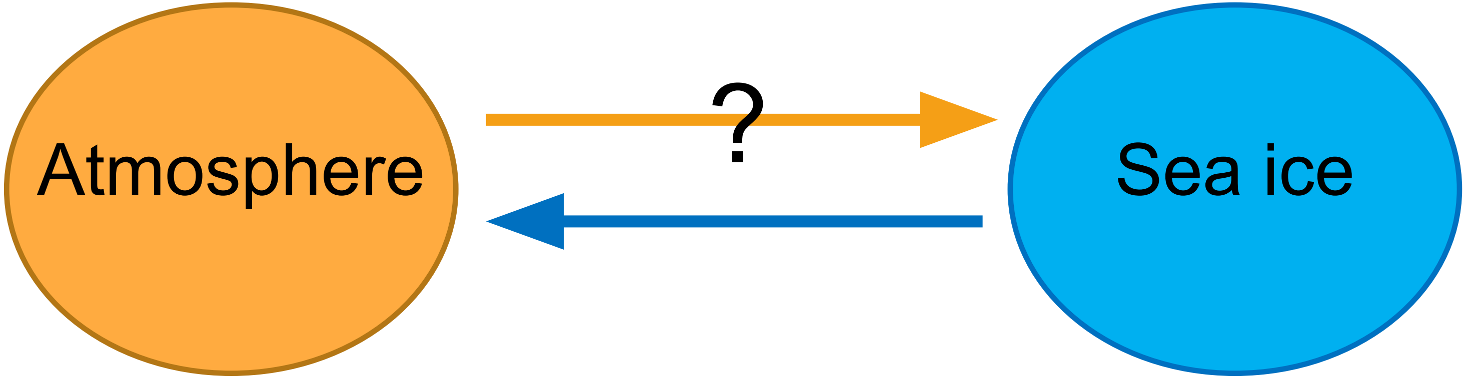
Arctic warming is almost twice as large as global average



Why are temperatures warming faster in the Arctic than the rest of the world?

Scientific questions

- Does the atmosphere primarily drive the sea ice variations or does sea ice dominate changes in atmosphere, over the Arctic?
- Are global climate models capable to capture this relationship?



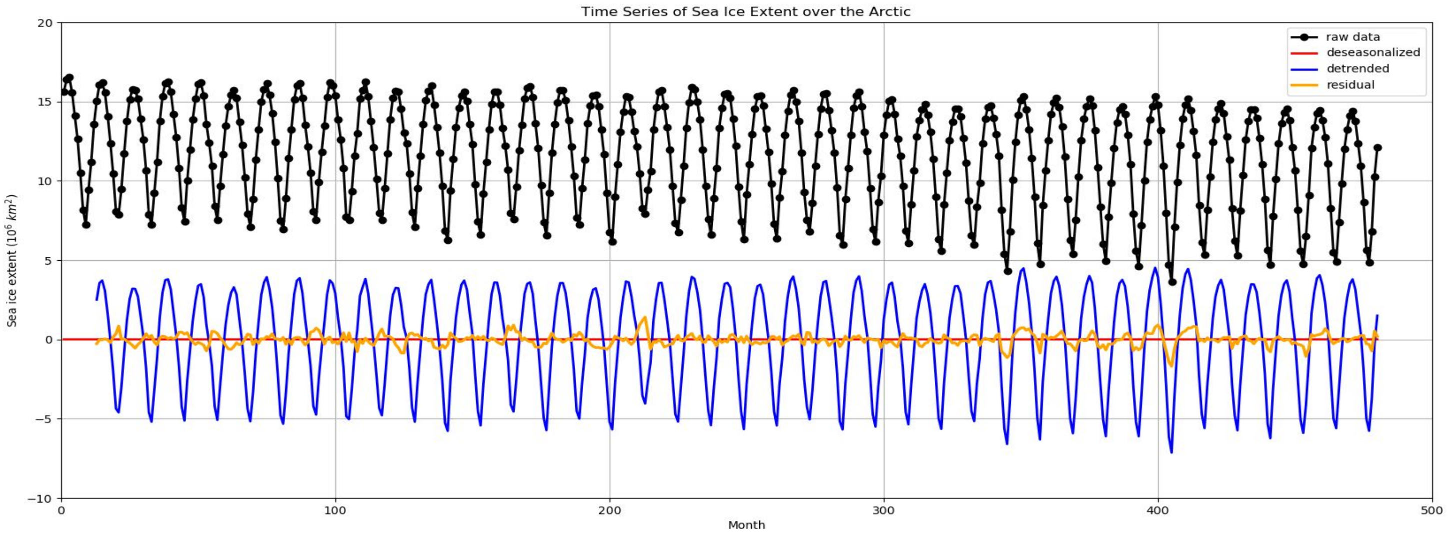
Data sets

Category	Variables	Data source	Data set	Temporal resolution coverage
Sea ice	Sea ice extent	National Snow and Ice Data Center/ National Aeronautics and Space Administration	Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS Passive Microwave Data, Version 1	11/1978-12/2018, Monthly
Thermodynamics	Air temperature	European Centre for Medium-Range Weather Forecasts	ERA-5 global reanalysis	01/1979-12/2019, Monthly
	Total precipitation	European Centre for Medium-Range Weather Forecasts	ERA-5 global reanalysis	01/1979-12/2019, Monthly
	Relative humidity	European Centre for Medium-Range Weather Forecasts	ERA-5 global reanalysis	01/1979-12/2019, Monthly
	Total cloud fraction, total cloud water path	European Centre for Medium-Range Weather Forecasts	ERA-5 global reanalysis	01/1979-12/2019, Monthly
	Surface sensible and latent heat flux, Surface downwelling shortwave flux, Surface downwelling longwave flux	European Centre for Medium-Range Weather Forecasts	ERA-5 global reanalysis	01/1979-12/2019, Monthly
Dynamics	Sea level pressure	European Centre for Medium-Range Weather Forecasts	ERA-5 global reanalysis	01/1979-12/2019, Monthly
	Geopotential heights at 850 hPa, 500 hPa and 200 hPa	European Centre for Medium-Range Weather Forecasts	ERA-5 global reanalysis	01/1979-12/2019, Monthly
	U wind, V wind and wind speed at 10 m	European Centre for Medium-Range Weather Forecasts	ERA-5 global reanalysis	01/1979-12/2019, Monthly

Pre-processing of data and other analysis

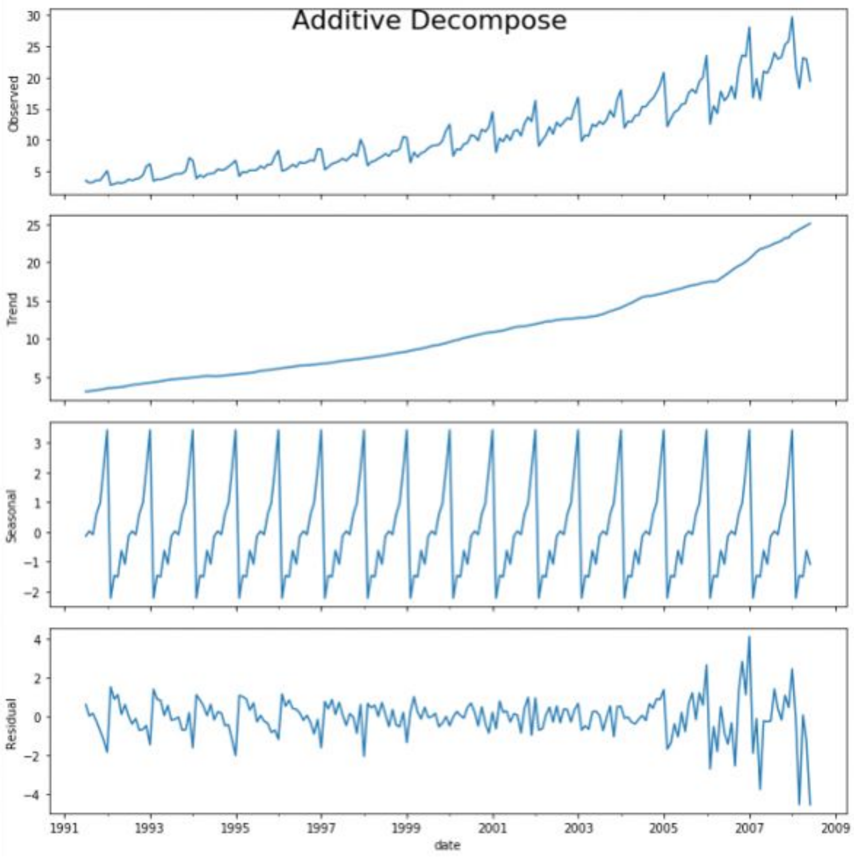
- Reduced the variables: Replaced GH_200hPa, GH_500hPa and GH_850hPa with their mean
- Normalized all the variables so that weights are not disproportionate
- Sensitivity Analysis of Hyperparameters
- Prepared a Causality Graph based on Domain Knowledge

Data pre-processing and time series decomposition



- Read gridded data (nc format) and average all data points within the Arctic domain (>60˚N)
- Create the time series (40 years x 12 months) for each variable and save it into CSV file
- Apply additive model to each variable to get the detrended, deseasonalized and residual components

Time series decomposition



Depending on the nature of the trend and seasonality, a time series can be modeled as an additive, wherein, each observation in the series can be expressed as a sum of the components:

The additive model is Y[t] = Trend[t] + Seasonality[t] + Residual[t]

- Detrend a time series

Subtract the line of best fit from the time series. The line of best fit was obtained from a linear regression model with the time steps as the predictor.

- Deseasonalize a time series

Divide the averaged seasonal index from the time series. The seasonal index were calculated from moving averages with 12-month seasonal window.

Lagging of variables for Temporal Graph

ONLY NEEDED for NOTEARS and DAG-GNN

- First convert time series X, Y, Z to variables X(t), X(t-1), X(t-2), Y(t), Y(t-1), Y(t-2), Z(t), Z(t-1), Z(t-2).

- Calculate causality graph among these variables. Then convert the graph to only have nodes X, Y and Z.

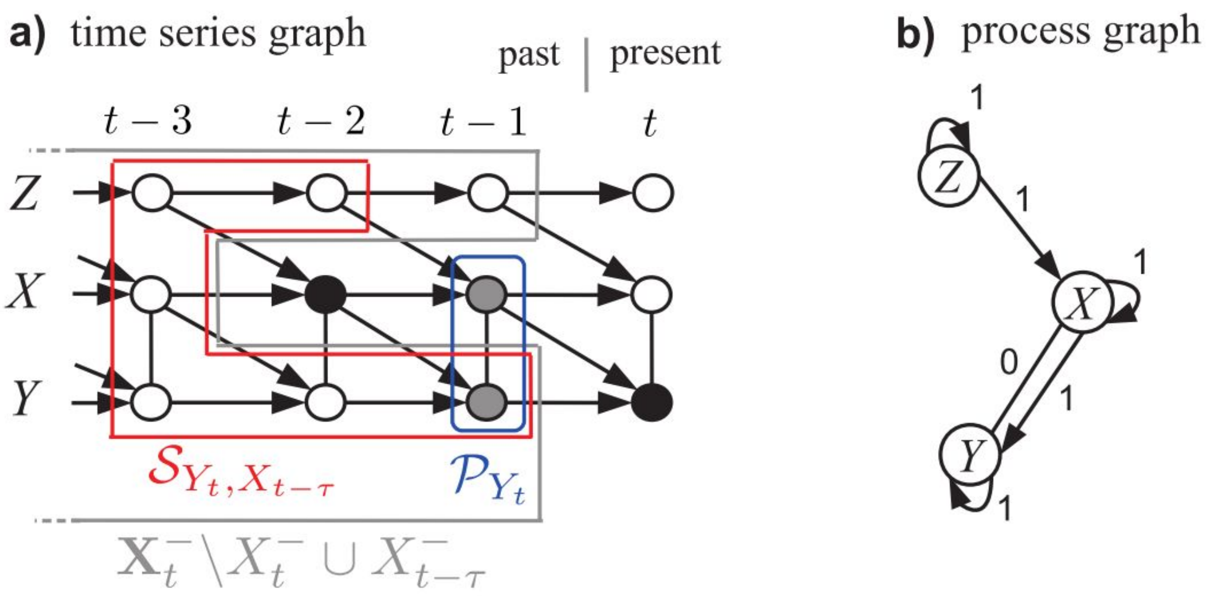
Similar idea to

- Figure 1 in “**Granger causality vs. dynamic Bayesian network inference: a comparative study**” C. Zhou and J. Feng. BMC Bioinformatics, 2009
- Figure 2 in “**Escaping the curse of dimensionality in estimating multivariate transfer entropy**” J. Runge, J. Heitzig, V. Petoukhov, J. Kurths.

Lagging of variables for Temporal Graph

ONLY NEEDED for NOTEARS and DAG-GNN

- Figure 2 in "**Escaping the curse of dimensionality in estimating multivariate transfer entropy**" J. Runge, J. Heitzig, V. Petoukhov, J. Kurths.



Processing lagged variables for Temporal Graph

ONLY NEEDED for NOTEARS and DAG-GNN

1. Discarded all non-valued rows. Eg: Only data including and below row 26 will be considered as training data.

1	6	Residual_heat_flux-7	Residual_heat_flux-8	Residual_heat_flux-9	Residual_heat_flux-10	Residual_heat_flux-11	Residual_heat_flux-12	Residual_shortwave-1	Residual_shortwave-2	Residual_shortwave-3	Resid
2											
3											
4											
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15								0.03407594447287470			
16								-0.27225121418814500	0.03407594447287470		
17								0.9566878035396730	-0.27225121418814500	0.03407594447287470	
18								1.4046526889905000	0.9566878035396730	-0.27225121418814500	0.034
19								-3.599604993358950	1.4046526889905000	0.9566878035396730	-0.272
20	060							3.6958395373197000	-3.599604993358950	1.4046526889905000	0.95
21	800	0.6591978652076060						-6.285752080427980	3.6958395373197000	-3.599604993358950	1.40
22	340	2.5668697458784800	0.6591978652076060					4.466985940400340	-6.285752080427980	3.6958395373197000	-3.5
23	510	0.7518546519930340	2.5668697458784800	0.6591978652076060				1.981920084989880	4.466985940400340	-6.285752080427980	3.69
24	130	-2.240062611839510	0.7518546519930340	2.5668697458784800	0.6591978652076060			-0.31755901056924100	1.981920084989880	4.466985940400340	-6.2
25	100	-0.6434214766435130	-2.240062611839510	0.7518546519930340	2.5668697458784800	0.6591978652076060		-0.2335687872691920	-0.31755901056924100	1.981920084989880	4.4
26	600	-0.056138518555528100	-0.6434214766435130	-2.240062611839510	0.7518546519930340	2.5668697458784800	0.6591978652076060	-0.1869714909795160	-0.2335687872691920	-0.31755901056924100	1.9
27	700	0.27682667879314600	-0.056138518555528100	-0.6434214766435130	-2.240062611839510	0.7518546519930340	2.5668697458784800	-0.3815046437565710	-0.1869714909795160	-0.2335687872691920	-0.317
28	300	-2.5210299176318700	0.27682667879314600	-0.056138518555528100	-0.6434214766435130	-2.240062611839510	0.7518546519930340	-0.207722566320669	-0.3815046437565710	-0.1869714909795160	-0.23
29	670	-1.5043563900035300	-2.5210299176318700	0.27682667879314600	-0.056138518555528100	-0.6434214766435130	-2.240062611839510	-0.4934039255218680	-0.207722566320669	-0.3815046437565710	-0.18
30	700	-0.2831625343161670	-1.5043563900035300	-2.5210299176318700	0.27682667879314600	-0.056138518555528100	-0.6434214766435130	-1.7655238675923100	-0.4934039255218680	-0.207722566320669	-0.38

2. Normalized each column from the resulting data

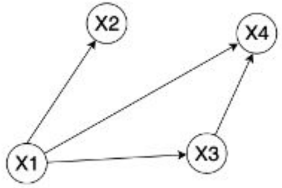
Causality/Causation/Cause and Effect Overview

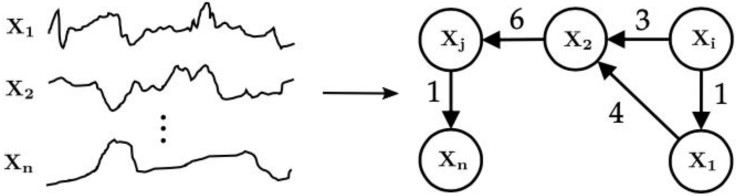
One process or state, a cause, contributes to the production of another process or state, an effect

The cause is partly responsible for the effect, and the effect is partly dependent on the cause

Examples:

$$
\begin{cases}
x_1(t) = 0.125 \cdot \sqrt{2} \cdot \exp(-x_1(t-1)^2/2) + \varepsilon_1 \\
x_2(t) = 1.2 \cdot \exp(-x_1(t-1)^2/2) + \varepsilon_2 \\
x_3(t) = -1.05 \cdot \exp(-x_1(t-1)^2/2) + \varepsilon_3 \\
x_4(t) = -1.15 \cdot \exp(-x_1(t-1)^2/2) \\
\quad + 0.2 \cdot \sqrt{2} \cdot \exp(-x_4(t-1)^2/2) \\
\quad + 1.35 \cdot \exp(-x_3(t-1)^2/2) + \varepsilon_4
\end{cases}
$$

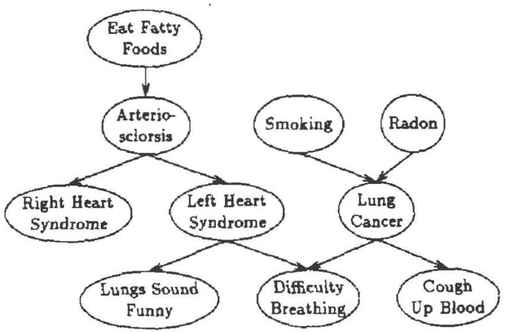




Causal Discovery Objectives

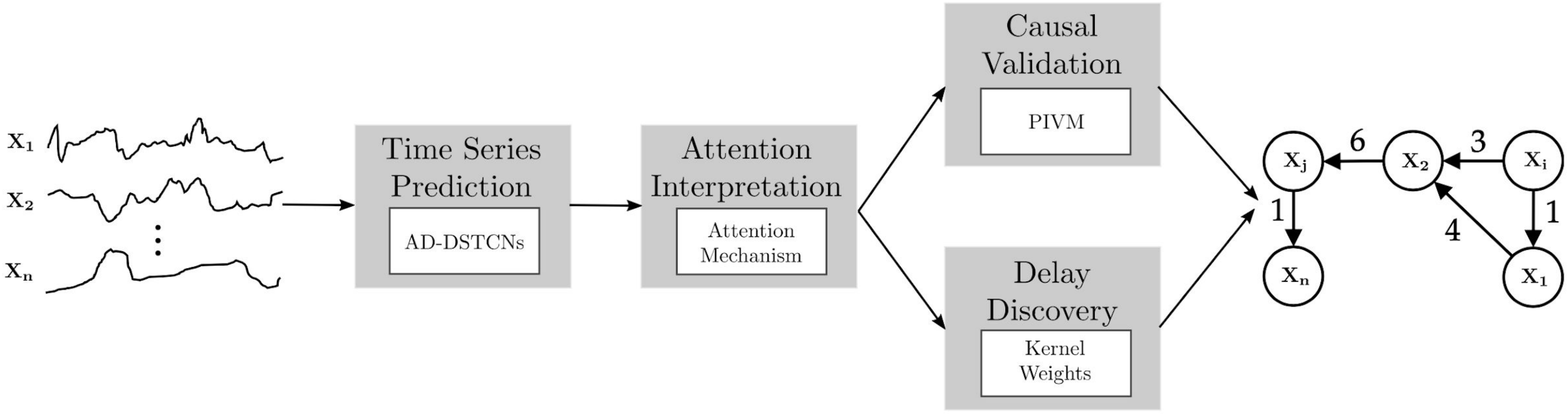
- Discover causal relationships between sea ice variations and atmospheric processes
- Using three state-of-the-art causal discovery methods
  1. TCDF
  2. NOTEARS algorithm
  3. DAG-GNN (builds upon NOTEARS)
- Visualize causal relationships through graphs

Example of Causal graph:



Method 1 for Causality Discovery (TCDF)

- Temporal Causal Discovery Framework (TCDF)
  - Attention-based CNN
  - Input: observational time series data
  - Output: Causality graph structure with time delay (lag)



Method 1 (TCDF) Causal Validation

A causal relationship is generally said to comply with two aspects:

1. Temporal precedence: the cause precedes its effect,
2. Physical influence: manipulation of the cause changes its effect.

To address:

1. Since the TCDF is temporal CNN, no info leakage from future to past.
2. Usually through interventions - keep all other variables value fixed, and change X_i to see the changes in X_j.
    a. Controlled experiments are hard to achieve
    b. Data-driven solutions: models the difference in evaluation score between original data and intervened dataset

Method 1 (TCDF) Permutation Importance (PI)

PI: measures how much an error score increases when the values of a variable are randomly permuted

Permuting a time series' values removes chronologicity and therefore breaks a potential causal relationship between cause and effect.

Only if the loss of a network increases significantly when a variable is permuted, the variable is a cause of the predicted variable.

Similar to Granger's causality validation: compare the loss of removing a variable

Method 2 (NOTEARS)

- Linear Structural Equation Model (SEM) with least-squares loss

$W \in \mathbb{R}^{d \times d}$ ... weighted adjacency matrix of graph G_W

Structure learning for linear *Structure Equation Model* (SEM):

$$\min_{W \in \mathbb{R}^{d \times d}} \|\boldsymbol{X} - \boldsymbol{X}W\|_F^2 + \lambda \|W\|_1 \quad \text{subject to} \quad G_W \text{ is a DAG} \qquad (1)$$

The paper shows that for a certain smooth function $h : \mathbb{R}^{d \times d} \to \mathbb{R}$

$$G_W \text{ is a DAG} \quad \Leftrightarrow \quad h(W) = 0$$

and proposes to solve (1) by solving

$$\min_{W \in \mathbb{R}^{d \times d}} \|\boldsymbol{X} - \boldsymbol{X}W\|_F^2 + \lambda \|W\|_1 \quad \text{subject to} \quad h(W) = 0 \qquad (2)$$

by means of the augmented Lagrangian method.

Method 3 (DAG-GNN)

- They learn the weighted adjacency matrix of a DAG by using a dee generative model that generalizes linear SEM
- In a way -- they are able to learn nonlinear SEMs, whereas NO TEARS paper was only learning linear SEMs

NO TEARS: $$X = (I - A^T)^{-1} Z.$$ Linear SEM

Here Z is the encoded latent variable of X

DAG-GNN: $$X = f_2((I - A^T)^{-1} f_1(Z)).$$ Non-linear SEM

Method 3 (DAG-GNN) Architecture and Loss Function

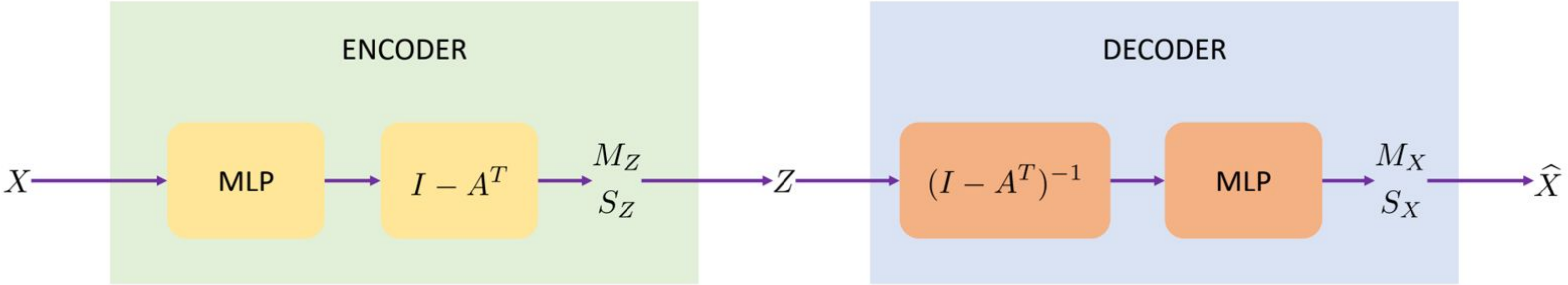


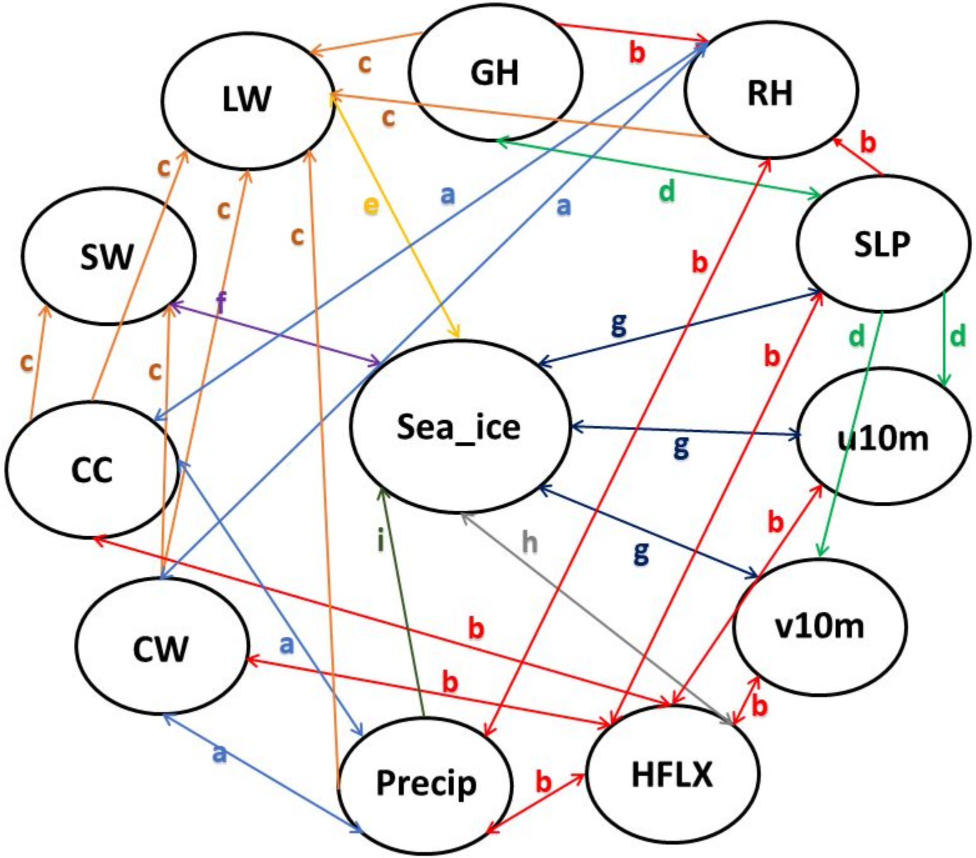
Figure 1. Architecture (for continuous variables). In the case of discrete variables, the decoder output is changed from M_X, S_X to P_X.

- Let $f_1 = 1$, i.e. identity mapping ; and f_2 = MLP.
- Nonlinear MLP better captures any nonlinearities than linear SEM (NOTEARS)
- Above (Figure 1) Architecture naturally handles discrete variables

Table of atmospheric and sea ice variables abbreviations used

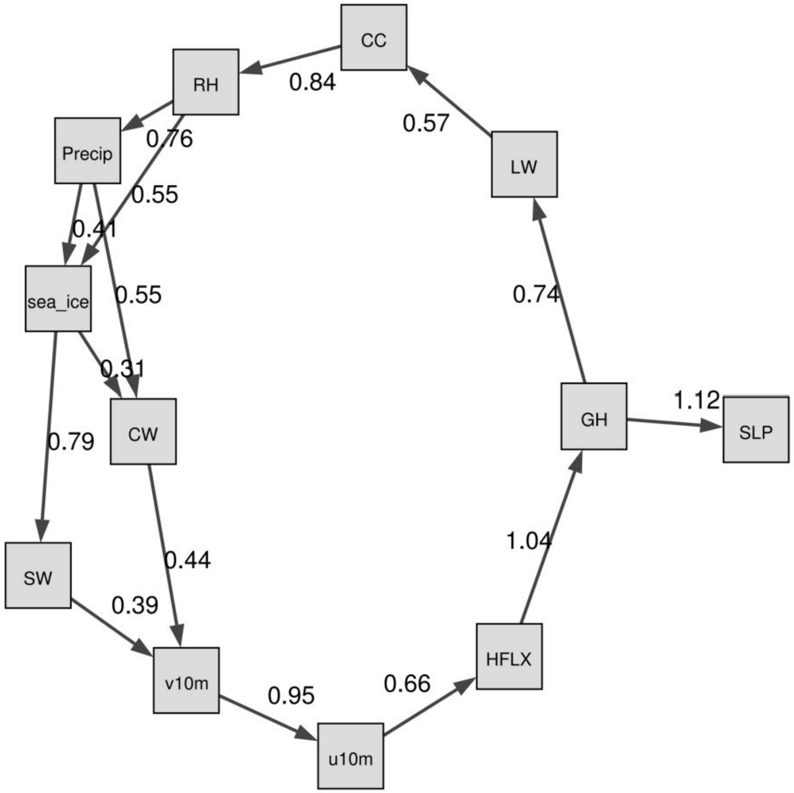
GH	Geopotential heights averaged from 200 hPa, 500 hPa and 850 hPa
RH	Relative humidity averaged from 1000-300 hPa
SLP	Sea level pressure
u10m	Zonal (u-component) wind at 10 meters
v10m	Meridional (v-component) wind at 10 meters
HFLX	Sensible and latent heat flux
Precip	Total precipitation
CC	Total cloud cover
CW	Total cloud water path
SW	Net shortwave flux at the surface
LW	Net longwave flux at the surface
Sea_ice	Sea ice extent in the Northern Hemisphere

Domain Knowledge graph

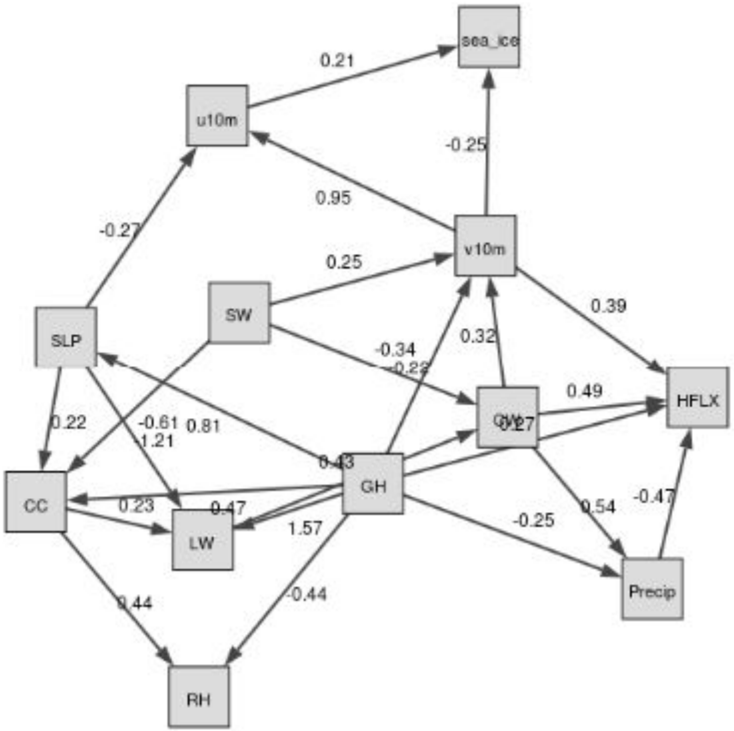


a. Cloud microphysics (e.g., Pruppacher and Klett, 1980, Nature)
b. Thermodynamics (e.g., Wallace and Hobbs , 2006, Elsevier)
c. Radiation (e.g., Liou et al. 2002, Elsevier)
d. Dynamics (e.g., Holton and Hakim, 2013, Academic press)
e. Kapsch et al. (2013, Nat. Clim. Change); Kapsch et al. (2019, Clim. Dyn.); Huang et al. (2017, JGR); Huang et al. (2019, GRL)
f. Kay et al. (2008, GRL); Choi et al. (2014, JGR); Kapsch et al. (2019, Clim. Dyn.)
g. Overland and Wang (2010, Tellus A); Watanabe et al. (2006, GRL); Wang et al. (2008); Rinke et al. (2019, JGR)
h. Boisvert et al. (2015, JGR; 2015, GRL); Bintanja and Selten (2014, Nature)
i. Perovich et al. (2002, JGR); Sturm et al. (2002, JGR);Boisvert et al. (2018, J. Clim.); Wang et al. (2019, Cryosphere)

Static model Results

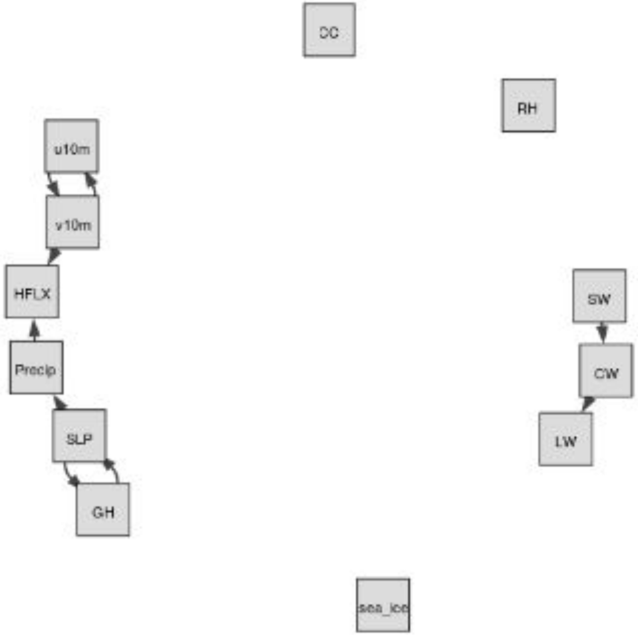


NOTEARS

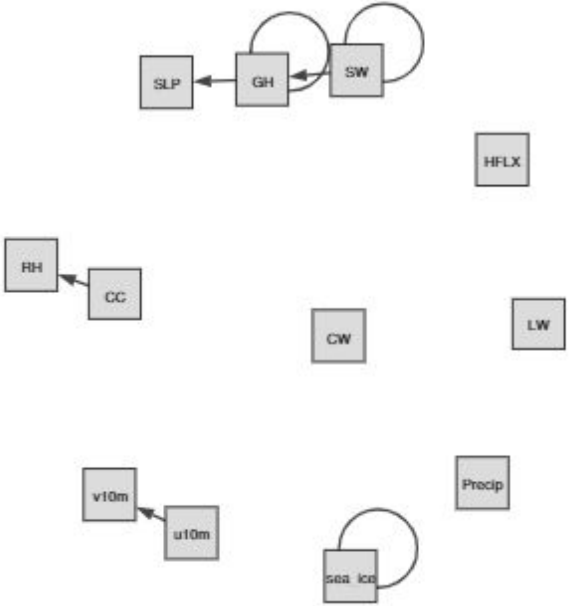


DAG-GNN

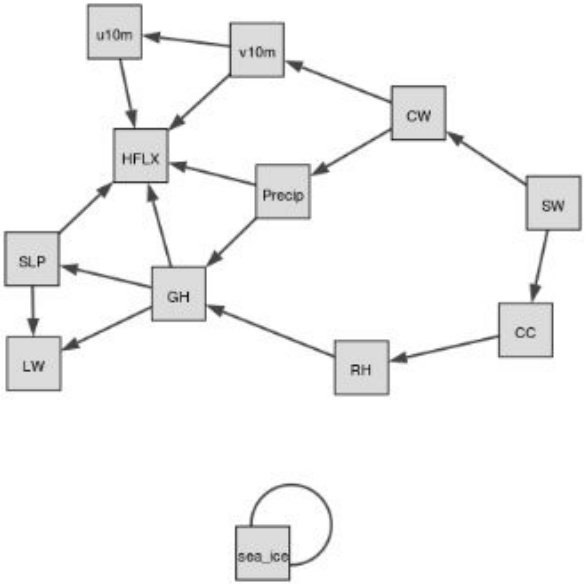
Temporal model Results



TCDF



NOTEARS



DAG-GNN

Sensitivity to Hyperparameters: TCDF

Table 5.1: Distance matrix with respect to the normalized Hamming distance for TCDF. ♣ denotes $layer = 0$, $kernel = 4$ are the algorithm's default hyperparameters. The bottom row compares to the domain knowledge graph of Figure 2.1 (best values in bold).

		Temporal					
		$layer = 0$ $kernel = 2$	$layer = 0$ $kernel = 4$♣	$layer = 0$ $kernel = 6$	$layer = 1$ $kernel = 2$	$layer = 1$ $kernel = 4$	$layer = 1$ $kernel = 6$
Temporal	$layer = 0, kernel = 2$	0	0.05	0.01	0.02	0.01	0.01
	$layer = 0, kernel = 4$♣	0.05	0	0.06	0.07	0.06	0.06
	$layer = 0, kernel = 6$	0.01	0.06	0	0.01	0.01	0.01
	$layer = 1, kernel = 2$	0.02	0.07	0.01	0	0.02	0.02
	$layer = 1, kernel = 4$	0.01	0.06	0.01	0.02	0	0
	$layer = 1, kernel = 6$	0.01	0.06	0.01	0.02	0	0
	Domain knowl.	0.35	**0.33**	0.34	0.34	**0.33**	**0.33**

Sensitivity to Hyperparameters: NOTEARS

Table 5.2: Distance matrix with respect to the normalized Hamming distance for NOTEARS. ♣ denotes that $\lambda = 0.1, t = 0.3$ are the algorithm's default hyperparameters. The bottom row compares to the domain knowledge graph of Figure 2.1 (best values in bold).

		Static				*Temporal*			
		$\lambda = 0$, $t = 0.2$	$\lambda = 0$, $t = 0.3$	$\lambda = 0.1$, $t = 0.2$	$\lambda = 0.1$, $t = 0.3$♣	$\lambda = 0$, $t = 0.2$	$\lambda = 0$, $t = 0.3$	$\lambda = 0.1$, $t = 0.2$	$\lambda = 0.1$, $t = 0.3$♣
Static	$\lambda = 0, t = 0.2$	0.0	0.02	0.15	0.15	0.54	0.36	0.16	0.15
	$\lambda = 0, t = 0.3$	0.02	0.0	0.15	0.12	0.53	0.35	0.14	0.12
	$\lambda = 0.1, t = 0.2$	0.15	0.15	0.0	0.02	0.51	0.36	0.09	0.1
	$\lambda = 0.1, t = 0.3$♣	0.15	0.12	0.02	0.0	0.52	0.35	0.07	0.08
Temporal	$\lambda = 0, t = 0.2$	0.54	0.53	0.51	0.52	0.0	0.18	0.48	0.51
	$\lambda = 0, t = 0.3$	0.36	0.35	0.36	0.35	0.18	0.0	0.33	0.34
	$\lambda = 0.1, t = 0.2$	0.16	0.14	0.09	0.07	0.48	0.33	0.0	0.03
	$\lambda = 0.1, t = 0.3$♣	0.15	0.12	0.1	0.08	0.51	0.34	0.03	0.0
	Domain knowl.	0.35	**0.33**	0.36	0.35	0.54	0.46	0.37	**0.35**

Sensitivity to Hyperparameters: NOTEARS

Table 5.4: Distance matrix with respect to the l_1-distance for NOTEARS. ♣ denotes that $\lambda = 0.1, t = 0.3$ are the algorithm's default hyperparameters.

		Static				*Temporal*			
		$\lambda = 0$ $t = 0.2$	$\lambda = 0$ $t = 0.3$	$\lambda = 0.1$ $t = 0.2$	$\lambda = 0.1$ $t = 0.3$♣	$\lambda = 0$ $t = 0.2$	$\lambda = 0$ $t = 0.3$	$\lambda = 0.1$ $t = 0.2$	$\lambda = 0.1$ $t = 0.3$♣
Static	$\lambda = 0, t = 0.2$	0.0	0.8	12.54	12.37	77.58	51.58	16.73	14.2
	$\lambda = 0, t = 0.3$	0.8	0.0	12.27	11.58	77.36	51.36	15.93	13.41
	$\lambda = 0.1, t = 0.2$	12.54	12.27	0.0	0.69	77.34	52.46	11.24	9.8
	$\lambda = 0.1, t = 0.3$♣	12.37	11.58	0.69	0.0	77.6	52.29	10.55	9.11
Temporal	$\lambda = 0, t = 0.2$	77.58	77.36	77.34	77.6	0.0	26.0	69.0	73.0
	$\lambda = 0, t = 0.3$	51.58	51.36	52.46	52.29	26.0	0.0	47.0	49.0
	$\lambda = 0.1, t = 0.2$	16.73	15.93	11.24	10.55	69.0	47.0	0.0	4.0
	$\lambda = 0.1, t = 0.3$♣	14.2	13.41	9.8	9.11	73.0	49.0	4.0	0.0

Sensitivity to Hyperparameters: DAG-GNN

Table 5.3: Distance matrix with respect to the normalized Hamming distance for DAG-GNN. ♣ denotes the algorithm's default hyperparameters. The bottom row compares to the domain knowledge graph of Figure 2.1 (best values in bold).

		Static				*Temporal*			
		$\tau=0$		$\tau=10^{-7}$		$\tau=0$		$\tau=10^{-7}$	
		$t=0.2$	$t=0.3$♣	$t=0.2$	$t=0.3$	$t=0.2$	$t=0.3$♣	$t=0.2$	$t=0.3$
Static	$\tau=0, t=0.2$	0.0	0.06	0.04	0.07	0.1	0.12	0.1	0.12
	$\tau=0, t=0.3$♣	0.06	0.0	0.05	0.01	0.08	0.07	0.08	0.07
	$\tau=10^{-7}, t=0.2$	0.04	0.05	0.0	0.06	0.08	0.1	0.08	0.1
	$\tau=10^{-7}, t=0.3$	0.07	0.01	0.06	0.0	0.08	0.08	0.08	0.08
Temporal	$\tau=0, t=0.2$	0.1	0.08	0.08	0.08	0.0	0.03	0.01	0.03
	$\tau=0, t=0.3$♣	0.12	0.07	0.1	0.08	0.03	0.0	0.05	0.0
	$\tau=10^{-7}, t=0.2$	0.1	0.08	0.08	0.08	0.01	0.05	0.0	0.05
	$\tau=10^{-7}, t=0.3$	0.12	0.07	0.1	0.08	0.03	0.0	0.05	0.0
	Domain knowl.	0.33	0.33	0.35	**0.32**	0.35	**0.34**	0.36	**0.34**

Sensitivity to Hyperparameters: DAG-GNN

Table 5.5: Distance matrix with respect to the l_1-distance for DAG-GNN. ♣ denotes the algorithm's default hyperparameters.

		Static				*Temporal*			
		$\tau = 0$		$\tau = 10^{-7}$		$\tau = 0$		$\tau = 10^{-7}$	
		$t = 0.2$	$t = 0.3$♣	$t = 0.2$	$t = 0.3$	$t = 0.2$	$t = 0.3$♣	$t = 0.2$	$t = 0.3$
Static	$\tau = 0, t = 0.2$	0.0	9.0	6.0	10.0	14.0	17.0	14.0	17.0
	$\tau = 0, t = 0.3$♣	9.0	0.0	7.0	1.0	11.0	10.0	11.0	10.0
	$\tau = 10^{-7}, t = 0.2$	6.0	7.0	0.0	8.0	12.0	15.0	12.0	15.0
	$\tau = 10^{-7}, t = 0.3$	10.0	1.0	8.0	0.0	12.0	11.0	12.0	11.0
Temporal	$\tau = 0, t = 0.2$	14.0	11.0	12.0	12.0	0.0	5.0	2.0	5.0
	$\tau = 0, t = 0.3$♣	17.0	10.0	15.0	11.0	5.0	0.0	7.0	0.0
	$\tau = 10^{-7}, t = 0.2$	14.0	11.0	12.0	12.0	2.0	7.0	0.0	7.0
	$\tau = 10^{-7}, t = 0.3$	17.0	10.0	15.0	11.0	5.0	0.0	7.0	0.0

Conclusions

- This study investigated the causality between multiple atmospheric processes and sea ice variations using three data-driven causality discovery approaches (TCDF, NOTEARS and DAG-GNN).
    - One advantage of utilizing these approaches is they not only generate causal graphs, but also provide quantified information on causal strength weight time lag.
    - We found that the outputs of the three algorithms are rather sensitive to the choice of hyperparameters.
        - Hence, some care must be taken when applying data-driven causality discovery approaches and domain knowledge is indispensable for assessing whether their produced outputs are reasonable.
    - Nevertheless, this is a pioneer study in the application of data-drive causality discovery approaches in the atmosphere-sea ice feedbacks.

References

- Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS:Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, 2018.
- Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG Structure Learning with Graph Neural Networks. In *International Conference on Machine Learning*, 2019.
- Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, January 2019.
- Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, 2020.