Comparative Evaluation of Causal Discovery and Inference Approaches on Arctic Sea Ice Time Series Data

Omar Faruque
Department of Information Systems
University of Maryland, Baltimore County (UMBC)
Maryland, USA
omarfaruque@umbc.edu

Xingyan Li
Department of Information Systems
UMBC
Maryland, USA
xingyanli@umbc.edu

Md Azim Khan
Department of Information Systems
UMBC
Maryland, USA
azimkhan22@umbc.edu

Homayra Alam
Department of Information Systems
UMBC
Maryland, USA
halam3@umbc.edu

Jianwu Wang
Department of Information Systems
UMBC
Maryland, USA
jianwu@umbc.edu

***Abstract*—Sea ice extent plays a crucial role in the Arctic system, and thus the study on causal relationships between sea ice extent and other climate variables comes to our interest to better understand the system. To find the causal relationship we applied various state-of-the-art causal discovery techniques from the time-independent and time-dependent domains. Then we employed several causal inference models to quantify the causal effects of different causal relationships in the Arctic system. The NSIDC Sea Ice Concentration observation data and the ERA-5 global reanalysis data were used in our study. Our analysis shows that the GES and VarLiNGAM from causal discovery methods and the conditional instrumental variable (CIV) causal inference model perform better on the Arctic Sea Ice time series dataset.**

***Index Terms*—Causal Discovery, Causal Inference, Time Series Data, Arctic Data Analysis**

I. Introduction

A. Background

The summer sea ice extent in the Arctic area reaches the minimum value this summer and has been shrinking each decade since 1979 at a rate of 12.2%, according to NASA Global Climate Change[1]. It is a warning because Arctic sea ice plays a vital role in circulating ocean water, regulating the air temperature, and preventing global warming. Therefore we would like to conduct a causality study to find what components in the Arctic climate system give rise to changes in sea ice extent and corresponding causal effects over time. This will help us understand relationships between different components of the Arctic climate system and their collective effect on the extent of sea ice.

[1]GLOBAL CLIMATE CHANGE - Vital Signals of the Planet. Arctic Sea Ice Minimum Extent. https://climate.nasa.gov/vital-signs/arctic-sea-ice/

This work is supported by the NSF grant "CAREER: Big Data Climate Causality" (OAC-1942714).

Causal discovery is the task of identifying causal relationships between variables in a complex system, and data-driven approaches have gradually taken the place of traditional methods such as randomized controlled trials [1]. Data-driven causal discovery methods can be roughly divided into Constraint-Based (CB) and Score-Based (SB) methods. The former category of algorithms satisfies the conditional independence test for a given dataset and includes approaches PC, FCI, RFCI, and CCD algorithms [2]. The latter fit data into goodness-of-fit tests by conducting a certain score criterion, and one typical model is GES [3]. On top of these approaches based on i.i.d. assumption, our study focuses more on causal discovery for time series data, which aims to learn interdependencies amongst instances in the sequence of data. In the past years, researchers proposed machine learning and deep learning-based causal discovery methods such as Granger Causality [4], VarLiNGAM [5], PCMCI [6], PCMCI+ [7], TCDF [8], NAVAR [9], TS-CausalCNN [10]. To be more specific, we will find causal relationships among variables with a time delay in impact to infer a causal graph using time-dependent causal discovery methods, and compare the results with domain knowledge-based causal graphs [11].

Another principal task is causal inference, which refers to quantifying the effect of a cause variable on a target variable based on the causal relationship found through causal discovery. Compared with traditional methods, advanced methods use neural networks such as RNN and LSTM [12]–[14] to learn latent representations of complex feature spaces. The advanced causal inference methods include TARnet [15], CEVAE [16], Time-Based Regression (TBR) [17], C-ARIMA [18], TCINet [19] and Counterfactual Recurrent Networks (CRN) [20]. Here we tend to focus more on time-dependent causal effects exerted on the "sea ice extent" variable. There

are two challenges in implementing causal inference: (1) how to define and modify treatment effect for time series and continuous data; (2) how to decide evaluation metrics for the sea ice dataset.

We aim to learn the existence and strength of causal relationships between time series components (atmosphere, ocean, sea ice sheet) in the Arctic region climate system and sea ice sheet. The research contributions are as follows. 1) Data pre-processing: convert the source dataset format and eliminate missing values and outliers. 2) Infer a causal graph from variables using causal discovery models. Compare results from time-dependent and IID methods. 3) Use machine causal inference techniques to estimate causal effects. 4) Evaluate the causal discovery and inference results with reasonable metrics.

B. Related Works

Climate models are critical to interpreting climate change. Causal discovery methods have a huge opportunity to apply to these models for better understanding. One of the most prominent computational approaches to interpreting causality is Granger Causality [21]. It first performs linear regression of the time series and then applies a statistical test on the regression coefficient. A good amount of research is done on structural learning for causal discovery. Chu et al. [22] applied structural learning in time series data of remote geospatial indices of ocean temperature to intercept the relationship between four ocean climate indices. Ebert-Uphoff et al. [22] discuss a brief introduction to several causal discovery algorithms using graphical models for climate science. Nichol et al. analyzed Data-driven causal evaluation of climate models for the Energy Exascale Earth System (E3SM) project [23]. They found differences in causal relationships between climate models and observed data using the PCMCI method. Deser et al. [24], Petoukhov and Semenov [25], Handorf et al. [26] approach to interpret climate models to investigate atmospheric changes using causal effects forced by input data. The Causal Effect Networks (CEN) method is analyzed for graphical models to overcome spurious relations due to autocorrelation, indirect effects, or common drivers of midlatitude winter circulation [27]. This work is extended by Samarasinghe using Granger Causality and Pearl Causality [28]. Granger Causality is used to validate the time sequence of the causal linkage for human society's impact on climate change [29]. They also analyze correlation and regression tests to legitimize the strength and consistency of causal linkages. Causal Model Evaluation (CME) is introduced to simulate atmospheric dynamical interactions between climate variables at remote locations [30]. Runge et al. used a nonlinear causal state space reconstruction method to extract the ecologically plausible causal network of sea surface temperatures [31].

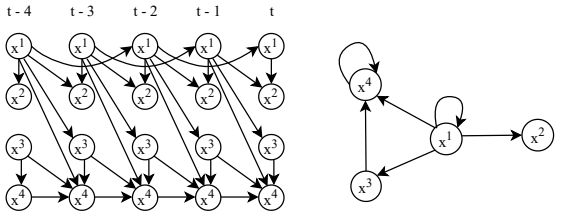
II. Preliminaries

Causal Discovery (CD) is the process of identifying cause and effect relationships between the variables of a system under study. CD approaches try to recover the underlying graph structure of the variables to represent how each variable interacts with other system variables. If the set of variables of a system is X, then CD can be expressed as $G(V, E) = CD(X)$ and will return a causal graph.

Causal Inference (CI) quantifies the effect of a cause variable on the target variable based on the causal relationship identified in the CD process. Another way of CI is to apply a treatment/intervention on the cause variable and measure the corresponding change in the effect variable considering all possible covariates. If it is not possible to get intervened data for any system, we can estimate the causal inference coefficient for a cause-effect relationship by considering covariates. Suppose $G(V, E)$ is a generated causal graph and $cause \rightarrow target$ is an edge. The causal effect from cause to target variable can be represented as $effect = CI(target, cause, covariates)$.

Full Causal Graph visualizes the causal relationship between variables across different time points showing exact time lags from cause to effect variable. It gives precise detailed information about interactions between variables of any dynamic system. An example of a full causal graph is given in Figure 1(a).

Summary Causal Graph is a simplified version of the full causal graph, showing causal links across variables without any temporal information. A summary causal graph only provides a high-level representation of the underlying causal structure. Figure 1(b) demonstrates an example of a summary causal graph.



a) Full Causal Graph b) Summary Causal Graph

Fig. 1. Example of full and summary causal graphs.

Counterfactual Data of a target variable is generated by applying a treatment to its cause variable and leaving other variables unchanged. For a population-based system, the initial population is divided into two subgroups, treatment is applied to one group to get the counterfactual data, and the group without applying treatment gives factual/observation data.

III. Dataset

The atmospheric and oceanic variables are collected across the Arctic Ocean region with daily records from 1979 to 2021. Variables in the dataset: (1) Sea ice extent collected from Nimbus-7 SSMR and DMSP SSM/I-SSMIS passive microwave data version[2]; (2) Nine other features are col-

[2]National Snow and Ice Data Center. Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS Passive Microwave Data, Version 1. http://nsidc.org/data/NSIDC-0051, 2021. Accessed: 2021-9-26.

TABLE I
NAME, SOURCE, AND UNIT OF EACH VARIABLE IN THE DATASET.

Variable	Source	Unit
Surface pressure	ERA5	Pa
Wind velocity (10 meter)	ERA5	m/s
Specific humidity	ERA5	kg/kg
Air temperature (2 meter)	ERA5	K
Shortwave radiation	ERA5	W/m^2
Longwave radiation	ERA5	W/m^2
Rainfall rate	ERA5	mm/day
Snowfall rate	ERA5	mm/day
Sea surface temperature	ERA5	K
Sea ice extent	NSIDC	Million km^2

lected from ERA-5 global reanalysis product[3] including wind velocity at 10-meter, specific humidity, longwave radiation, shortwave radiation, rainfall rate, snowfall rate, sea surface temperature, 2 meter air temperature, and sea level pressure (surface pressure). Table I shows the variable names and units.

The original dataset is a sequence of geographical 2-D data in NetCDF file format and pre-processed before fitting into different methods. The dimension of data is reduced by averaging pixel values across the Arctic region down to a single value for the whole region. In this way, we have one sequence of daily records for each variable respectively. The missing value of x_t is replaced with an average of x_{t-1} and x_{t+1} for all variables. Then for each month average of daily values is taken to have a dataset in monthly frequency.

IV. METHODS

This section describes relevant methods of causal discovery, causal inference, and evaluation criteria.

A. Causal Discovery Methods

Depending on the assumption of data distribution, causal discovery methods can be grouped into IID and time series-based CD methods. IID data focused CD methods are proposed for independent and identically distributed datasets, but we can apply these methods for time series data without considering temporal features.

1) CD Methods for IID Data:

a. PC algorithm

PC algorithm [32] is one of the earliest causal discovery algorithms, which uses the idea that two statistically independent variables are not causally linked. The PC algorithm starts with a complete undirected graph and then identifies the skeleton using conditional independence tests. If the conditional independence decisions are correct in the large sample limit, the PC algorithm is guaranteed to converge to the true Markov Equivalence Class, assuming the Causal Markov and Faithfulness assumptions and no unmeasured confounders.

b. Greedy Equivalence Search (GES)

GES [33] is a way to navigate a search space such that you always move in a direction that seems beneficial based on local surroundings. It's like being lost in a forest and trying to get out by only moving toward open areas. GES starts with an empty graph and iteratively adds directed edges to improve a model fitness measure. At each step in the algorithm, the edge that most improves the fit score is added to the graph. The resulting model is then mapped to the corresponding Markov equivalence class, and the procedure continues. When the score can no longer be improved, the algorithm then asks, edge by edge, which edge removal, if any, will most improve the score, until no further edges can be removed. The GES algorithm depends on the relative strength of associations and conditional associations of variables. In the large sample limit, two algorithms converge on the same Markov Equivalence Class. However, on finite samples, two algorithms may give different results.

c. Fast Causal Inference (FCI)

The FCI algorithm [34] tolerates and sometimes discovers unknown confounding variables. FCI constructs a causal graph starting with a fully connected undirected graph and removes edges that connect conditionally independent variables. The FCI algorithm is explained in the figure below. In the figure, the top subgraph (A) denotes the ground truth graph with a confounder that works as a common cause of Y and Z. After applying conditional independence all confounder edges will be removed. Finally, the bidirectional edge in output graph C explains that there is at least one unmeasured confounder of Y and Z.

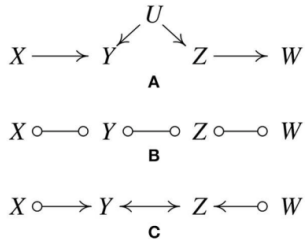


Fig. 2. FCI algorithm steps to find the existence of latent confounders [34].

2) CD Methods for Time Series Data:

a. PCMCI

PCMCI [6] is a causal discovery method for high-dimensional time series data featuring linear/non-linear, and time delayed dependencies. PCMCI is constructed based on the conditional independence framework, with assumptions of unconfoundedness, causal Markov condition, faithfulness, no contemporaneous causal effects, and stationarity. The algorithm deals with the trade-off between high false-positive rates and low true-positive rates in results using two stages: PC_1 condition selection based on the PC-stable algorithm and the momentary conditional independence (MCI) test. The PC_1 stage uses an iterative approach to estimate optimal parents $\tilde{P}(X_t^j)$ of variable X_t^j such that $\{X_t^j \perp \mathbf{X_t^-} \backslash \tilde{P}(X_t^j) | P(X_t^j)\}$, and the second stage test $\{X_{t-\tau}^i \perp X_t^j | \tilde{P}(X_t^j), \tilde{P}(X_{t-\tau}^i)\}$ (where τ is the maximum time delay) to prevent spurious edges caused by confounders. To combine both stages, three

[3]European Centre for Medium-Range Weather Forecasts. ERA-5 global reanalysis product. https://cds.climate.copernicus.eu/cdsapp!/home,2021. Accessed: 2021-9-25.

independence test methods are available including linear partial correlation (ParCorr), GPDC, and CMI. PCMCI method performs better compared with other earlier algorithms for large numbers of variables. The limitation of PCMCI is that state-space methods outperform PCMCI for highly deterministic systems.

b. PCMCI+

PCMCI+ [7] improves PCMCI for auto-correlated time series data by changing the conditional independence check. The model conditioned out noise in lagged adjacencies to remove more edges and yield higher recall. There are four phases in this algorithm: skeleton discovery phase with lagged conditions, skeleton discovery phase with contemporaneous conditions and MCI, collider orientation phase, and rule orientation phase. The first phase is similar to PCMCI to estimate the optimal set of lagged parents $\hat{B}_t^-$. The second phase removes all adjacencies using $\{X_{t-\tau}^i \perp X_t^j | \mathcal{S}, \hat{B}_t^-(X_t^j), \hat{B}_{t-\tau}^-(X_{t-\tau}^i\}$, where $\mathcal{S} = \mathbf{X_t^-} \backslash \{X_{t-\tau^i}\}$. The third phase orients contemporaneous links based on unshielded triples based on MCI tests. The fourth phase orients the remaining contemporaneous links. The results show higher recall and shorter computation time than PCMCI. Implementation of PCMCI and PCMCI+ are available in TIGRAMITE tool: https://github.com/jakobrunge/tigramite.

c. TCDF: Temporal Causal Discovery Framework

TCDF [8] is a deep learning framework to discover causal relationships between time series and generate a causal graph. TCDF uses Attention-based Convolutional Neural Networks combined with a causal validation step. By interpreting the internal parameters of the convolutional networks, TCDF can also discover the time delay between a cause and its effect. The learned temporal causal graphs can include confounders and instantaneous effects. This broadly applicable framework can be used to gain novel insights from causal dependencies of a complex system for reliable predictions, knowledge discovery, and data-driven decision making.

d. VarLiNGAM: Vector Auto-regressive Linear Non-Gaussian Acyclic Model

VarLiNGAM [35] is a Structural Equation Models (SEM) based method for causal discovery. It combines the basic LiNGAM model with the classic vector autoregressive models (VAR). It enables analyzing lagged and contemporaneous (instantaneous) causal relations of time series data. The effect time series $x(t)$ is estimated as $x(t) = \sum_{T=0}^{k} B_T x(t-T) + e(t)$, where t is the time index, T is the time lag, e is non-Gaussian external influences, B_T is matrix modeling effects with time lag T. The element b_{ij} at i-th row and j-th column of B_T represents the causal effect of j-th variable on i-th variable with time delay T.

e. Dimensional Causality (DC)

It is required to quantify the causal relationship between different variables of a dynamical system to explain its operation clearly. Different time series of a dynamical system can be represented and estimated with the time-lag embedding of that time series. According to Taken's theorem [36], using dimensions of the time-lagged embedded manifold of any time series is it possible to estimate the actual dynamics of a complex system, and the dimension of a dynamical system will be equivalent to the dimension of that time series in embedded space. Based on this theorem, the DC method [37] computes the dimension of cause (D_X) and effect (D_Y) variables in the time-lagged embedding. Joint dimension D_J of cause and effect is generated by multiplying the two embedded spaces and measuring the dimension of the resultant manifold. These dimensions can be used to estimate the direction of causality and the probability of causal effects between two variables (X and Y) using the following equations.

$$X \to Y \iff D_X < D_Y = D_J \tag{1}$$

$$X \perp Y \iff D_X + D_Y = D_J \tag{2}$$

$$X \leftrightarrow Y \iff D_X = D_Y = D_J \tag{3}$$

$$X \curlyvee Y \iff max(D_X, D_Y) < D_J < D_X + D_Y \tag{4}$$

B. Causal Inference Methods

1) CI Methods for IID Data:

a. Causal Effect Variational Autoencoder (CEVAE)

The inference of causal relationships between different treatments and outcomes from observational data mainly depends on measuring the confounders. Confounders have an effect on treatment and outcome at the same time. For measurable confounders, we can easily control their impact on treatment and outcome variables by adjusting these confounders. There is no established way to manage the effect of a confounder when it is not observable. This problem can be solved using observed potential confounders to estimate the unobserved confounders [38]. These observed potential confounders are also called proxy variables. For instance, in the graph below, t represents treatment, y represents result, Z represents unobserved confounder, and X represents observed proxy. So, by adjusting values of X we can control the effect of confounder Z on treatment and outcome.

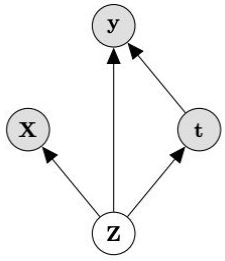


Fig. 3. Causal graph with a proxy variable X and unobserved confounder Z (t = treatment and y = outcome) [38].

However, finding the relationship between proxies and unobserved confounders is not trivial. Using different available proxies, the CEVAE [38] proposes a maximum-likelihood-based method to estimate a latent variable to discover the relationship between proxies and unobserved confounders, and the effect of these confounders on treatment and outcome. Using variational autoencoders (VAEs) based on the optimization of the variational lower bound likelihood, CEVAE learns the

latent variable causal model to infer the complex non-linear relationships between X and (Z, t, y) and roughly recovers $p(Z, X, t, y)$. After learning latent variables, the observed dataset is factorized by putting conditions on latent variables, and then the effect of treatment on outcome is measured. Through experimental results, it was demonstrated the CEVAE model achieves the best estimates of treatment effect and outperforms other state-of-the-art models such as BLR, BART, BNN, and CFRW.

b. Dragonnet

Dragonnet [39] uses a neural network to estimate the treatment effects using observational data in two steps. In the first stage, the model is fitted to estimate the propensity score across each unit according to outcome and covariates. For the second part, this fitted model is plugged into the downstream estimator. Dragonnet [39] comprised a three-headed architecture with an end-to-end procedure to estimate the propensity score with conditional outcomes from covariates and information on treatment. A deep neural network is used to learn a representation of X, $Z(X)\epsilon\mathbb{R}$ and for predicting both the treatment and outcome from this shared representation.

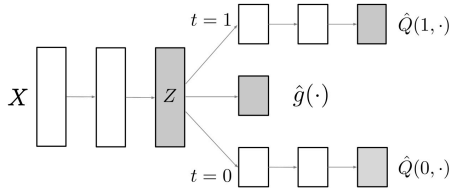


Fig. 4. Architecture of Dragonnet model [39].

The propensity score model ĝ(.) forces the representation layer to tightly couple to the estimated propensity scores. Using end-to-end training and high capacity, Dragonnet prevents from tossing away information. The trade-off is the quality of prediction and achieving a good representation of the propensity score. However, using a downstream estimator boosts up ATE estimation that does not use the estimated propensity scores.

c. T-Learner

Since one potential outcome is unobserved at a time, therefore supervised models are not applied directly to learn the effect of treatment. Instead, using machine learning to model each potential outcome individually and using plug-in estimators for treatment effects has become a popular strategy in econometrics, biostatistics, and machine learning. T-learner [40] method trains two independent models to learn potential outcomes instead of using a single neural network. Estimators for both models would be a feed-forward network with optimized MSE for predicting the observed outcomes. The joint loss function for a T-learner can be written as $L(Y; h(X; T)) = MSE(Y; h(X; 0)) + MSE(Y; h(X; 1))$. After the completion of training, both networks of T-learner have the same amount of input to predict the potential outcomes Ŷ(T) and Ŷ(1-T). These predictions are used to estimate the CATE for each unit,

$$\hat{T}_i = (1 - 2T_i)(\hat{Y}_i(T_i) - \hat{Y}_i(1 - T + i))$$

and the average treatment effect as,

$$\hat{ATE} = 1/N \sum_{i=1}^{N} \hat{T}_i$$

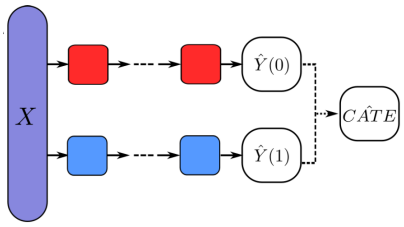


Fig. 5. In a T-learner, separate feed-forward networks are used to model each outcome [40].

2) CI Methods for Time Series Data:

a. Counterfactual recurrent Network (CRN)

CRN [20] causal inference approach is applied to recommend treatment medication and treatment timing for patients over time. The authors proposed a treatment invariant representation (balanced representation) by removing the association between patient history and treatment, thus removing bias caused by time-dependent confounders. The architecture of CRN includes an encoder and a decode. The encoder is an RNN network with LSTM units that predict one-step-ahead outcomes. The decoder uses balanced representation from the encoder to estimate counterfactual outcomes for future time steps. Assume X_t is time-variant confounders, Y_t is outcomes given treatment A_{t-1}, V is static features (e.g. genders), and H is patient history $H_t = (X_t, A_{t-1}, V)$, then with a maximum time lag τ, the output of CRN is $\mathbb{E}(\mathbf{Y}_{t+\tau}[\mathbf{a}(t, t + \tau - 1)]|(H)_t)$.

b. Variable-lag Granger Causality and Transfer Entropy

Granger causality is a fundamental method for generating causal inferences in time series data. The Granger causality method strongly assumes that each data point in a target variable is caused/affected by a series of other data points with a constant time lag. X granger causes Y, if the historical data of X predicts the value of Y more accurately than the historical data of Y does on its own. The transfer entropy method is a nonlinear extension of the Granger causality method that also takes into account the fixed time delay assumption. However, the assumption of fixed-lag cause is not valid for many real-world natural and social phenomena. For example, multiple initiators derived from time-series data with different time lags can affect stock markets, brain activity, environmental issues, etc. Variable-lag Granger causality and Variable-lag Transfer Entropy [41] method resolved this limitation. Dynamic Time Wrapping (DTW) is incorporated into Granger causality to measure the variable lag causes of X to Y. If X and Y be time series and $\delta max \in N$ be a maximum

time lag, the variable-lag granger causal relation is defied as $r^*_{YX}(t) = Y(t) - \sum_{i=1}^{\delta max}(a_i Y(t-i) + b_i X(t-i) + c_i X^*(t-i))$ and $X^*(t-i) = X(t-i+1-\Delta_{t-i+1})$, where $\Delta_t > 0$ is a time delay that minimizes the relation. X vl-granger causes Y if the variance of r^*_{YX} is less than the variance of both r_Y and r_{YX}. Similarly, for non-linear time series data the variable-lag transfer entropy is defined for two time series X and Y as follows: $\tau_{X \to Y} = H(Y(t)|Y^{(k)}_{t-1}) - H(Y(t)|Y^{(k)}_{t-1}, X^{(l)}_{t-1})$, where $H(.|.)$ is a conditional entropy, k and l are time lag, $Y^{(k)}_{t-1} = Y(t-1), ..., Y(t-k)$, and $X^{(l)}_{t-1} = X(t-1), ..., X(t-l)$. This method determines whether X variable-lag causes Y, X fixed-lag causes Y, or there is no conclusion of causation between X and Y using variable-lag Granger causality or transfer entropy given the maximum time lag δmax.

c. Time Series Deconfounder

Consider a treatment and effect relationship in a time step, the effect can become a cause over time and create complexity in future effect estimation. The Time Series Deconfounder [42] method uses a recurrent neural network architecture to enable the estimation of treatment effect leveraging the sequential assignment of multiple treatments over time in the presence of hidden confounders. This method creates a factor model to depict how the causes (treatments) are distributed over time. The latent variable $Z_t = g(\bar{h}_{t-1})$ is generated by the factor model at time t, where $\bar{h}_{t-1} = (\bar{a}_{t-1}, \bar{X}_{t-1}, \bar{Z}_{t-1})$ is the realization of history $\bar{H}_{t-1}$. The recurrent part of the factor model derives the latent variable Z_t in a way that depends on historical data. The outcome model calculates individual treatment impact over time based on latent variables, $E[Y|\bar{a}_{>=t}, \bar{A}_{t-1}, \bar{X}_t, \hat{Z}_t] = E[Y(\bar{a}_{>=t})|\bar{A}_{t-1}, \bar{X}_t, \hat{Z}_t]$. Gradient descent-based techniques can be used to train the factor model using treatment assignments from observational datasets.

f. Conditional Instrumental Variable (CIV)

In a causal graph, some variables are called instrumental variables (IV) if related to the cause, related to the effect only through the cause, and independent of the hidden confounders. Suppose a model $Y = \beta X + g(H, \varepsilon^Y)$ then using IVs it is possible to estimate coefficient β of the model [43]. However, observed time series in real dynamical systems show memory effect, which means the value of the current time also contains some portion of previous values of that time series. Consider the full causal graph in Figure 1(a). In Figure 6, target variable Y is dependent on instrument variable I through B. This leads us to introduce another set of conditioning variables B that make variable I and effect Y independent when we ignore all edges from $X \to Y$. None of the variables in the set B are descendants of set $(X \cup Y)$. Considering this additional conditioning set B with the instrumental variable I to measure the causal effect of X to Y is called the conditional instrumental variable. Using the CIV [43] method we can measure the value of the effect variable based on the cause and the coefficient β. Also, the residual of the estimated effect and ground truth effect is independent of I and B, $Y - \beta X \perp I|B$.

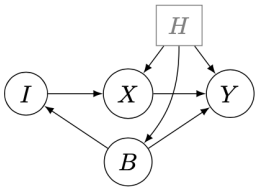


Fig. 6. Example of Conditional Instrumental Variable (CIV) [43].

C. Evaluation Metrics

1) Causal Discovery: We will evaluate causal discovery results by comparing generated causal graphs and the domain knowledge-based causal graph mentioned in [11]. The following evaluation metrics will be used to compare causal discovery results.

- True Positive Rate (TPR): is the probability that an actual positive will test positive. Assuming the threshold t of probability of an edge $p(a_{ij})$ ranges from $(0, 1)$, TPR is defined as $TPR_t = |\frac{\{(i,j):p(a_{ij}) \geq t\} \cap S}{S}|$ where S is the set of ground truth edges (i.e., $(i, j) : a^*_{ij} = 1$).
- False positive rate (FPR): is the probability that an actual negative will test positive. Assuming the threshold t of the probability of an edge $p(a_{ij})$ ranges from $(0, 1)$, FPR is defined as $FPR_t = |\frac{\{(i,j):p(a_{ij}) \geq t\} \cap \hat{S}}{\hat{S}}|$ where $\hat{S}$ is the set of ground truth missing edges (i.e., $(i, j) : a^*_{ij} = 0$).
- Structural Hamming Distance (SHD) [44]: is the number of operations required to make two causal graphs equivalent. The operations include adding, deleting an undirected edge, and adding, removing, or reversing the orientation of an edge.

However, since we lack a baseline model or causal diagram to evaluate our results, we may combine the causal graph driven by domain knowledge to conclude edges feasible for causal inference.

2) Causal Inference: The evaluation metrics best describe the quality of causal inference as follows.

- Precision in Estimation of Heterogeneous Effect (PEHE): The PEHE method is used to measure the accuracy of the individual treatment effect (ITE) estimation. It reflects the ability to capture individual variation in treatment effect estimation for a population. $PEHE = \frac{1}{N}\sum_{i=1}^{N}((y_{i1} - y_{i0}) - (\hat{y}_{i1} - \hat{y}_{i0}))^2$, where y_{i1} = the true outcome for $t = 1$, y_{i0} = the true outcome for $t = 0$, $\hat{y}_{i1}$ = the estimated outcome for $t = 1$, and $\hat{y}_{i0}$ = the estimated outcome for $t = 0$.
- Absolute error on the Average Treatment Effect (ATE): To measure the quality of the treatment effect estimation for the population level we can use the absolute error in the observed average treatment effect. The formula to compute the ATE absolute error is $\epsilon_{ATE} = |\frac{1}{n}\sum_{i=1}^{n}(y_i(1) - y_i(0)) - \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i(1) - \hat{y}_i(0))|$, where y_i represents the true treatment effect and $\hat{y}_i$ represents the estimated treatment effect.

V. Results

A. Causal Discovery Results

a. PC, GES, and FCI

We are analyzing the Arctic time series dataset but these methods expect IID data. We disregarded the time features and considered each observation independently distributed to apply these techniques to the Arctic dataset. Figure 7 shows the causal graph generated by the PC method. From the causal graph, we can see that air temperature and rainfall are the direct causes of sea ice extent, whereas, specific humidity is the effects of sea ice extent. Another important inference is between sea surface temperature and sea ice extent where no unidirectional causal relationship exists.

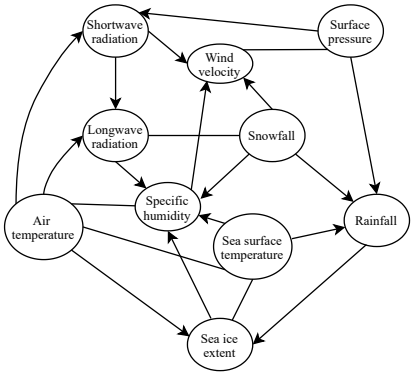


Fig. 7. Causal graph generated using the PC algorithm.

The causal graph generated using the GES method is shown in Figure 8. Here sea surface temperature, wind velocity, shortwave radiation, and air temperature are direct causes of sea ice extent. However, sea ice extent is a common cause of surface pressure and longwave radiation.

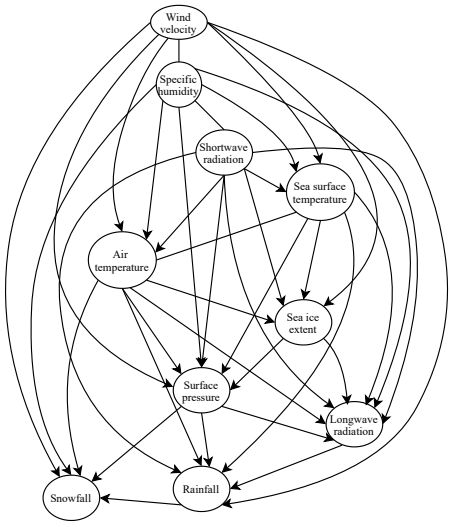


Fig. 8. Causal graph generated using the GES algorithm.

Figure 9 represents a causal graph generated by the FCI algorithm. As seen from the graph, sea surface temperature is the direct cause of sea ice extent. But there is an unmeasured confounder between specific humidity and sea ice extent, rainfall and sea ice extent, air temperature and sea ice extent.

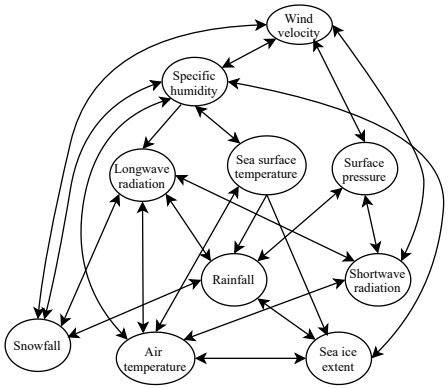


Fig. 9. Causal discovery results from the FCI algorithm.

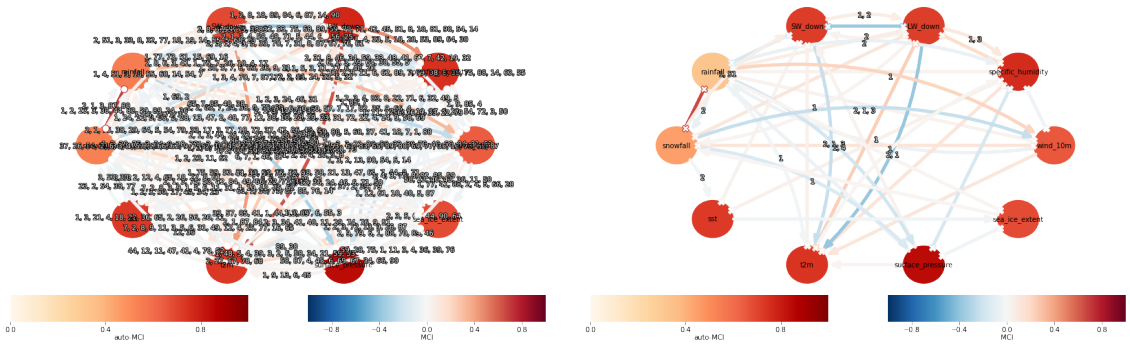
b. PCMCI, PCMCI+ and VarLiNGAM

The required input hyper-parameters for PCMCI and PCMCI+ include the maximum time delay τ_{max}, verbosity α to decide the detection power (high values give rise to high false positives and small values cause high dimensionality of the condition set), and independent test methods `cond_ind_test`. The following are the values used in this study: $\tau_{max} = 90$, $\alpha = 1$ (no parents are removed), `cond_ind_test = ParCorr`. Figure 10 (a) and (b) are causal graphs inferred by PCMCI and PCMCI+, respectively. VarLiNGAM also uses $\tau_{max} = 90$ as the maximum time delay, and it finds 1191 edges in total with causal effects larger than 0. Table II shows the top 15 edges related to sea ice extent found by VarLiNGAM.

TABLE II
CAUSAL EDGES RELATED TO SEA ICE EXTENT IDENTIFIED BY THE VARLINGAM METHOD.

Cause	Lag	Effect	Effect Coefficient
Snowfall	0	Sea ice extent	28120
Snowfall	3	Sea ice extent	15068
Snowfall	9	Sea ice extent	13920
Snowfall	17	Sea ice extent	12305
Snowfall	2	Sea ice extent	9736
Snowfall	80	Sea ice extent	8825
Snowfall	84	Sea ice extent	7426
Snowfall	45	Sea ice extent	7422
Snowfall	70	Sea ice extent	6314
Snowfall	74	Sea ice extent	5951
Rainfall	56	Sea ice extent	5541
Snowfall	26	Sea ice extent	5439
Snowfall	68	Sea ice extent	4897
Rainfall	52	Sea ice extent	4765
Snowfall	64	Sea ice extent	4762

Since we are interested in the change of sea ice extent, we look into the relationship between sea ice extent and other variables. From the causal graphs, if we only inspect



(a) (b)

Fig. 10. The causal graph inferred from (a) PCMCI and (b) PCMCI+.

the direct cause-effect edge type, we can see that for sea ice extent, **PCMCI** shows sea ice extent with time delay 1 has the strongest causal effect on itself, and rainfall (lag 0), snowfall (lag 0), wind velocity (lag 1), surface pressure (lag 2), sea surface temperature (sst) (lag 0) are the second strongest causes. **PCMCI+** shows that sea ice extent (lag 1) has the strongest causal effect on itself, and sst (lag 0) has the second strongest effect; meanwhile, sea ice extent (lag 0) can be a cause of shortwave radiation (SW_down). **VarLiNGAM**, interestingly, shows that snowfall (lag 0) has the strongest causal effect on sea ice extent, and rainfall (lag 56) is the second. According to a summary of the findings from these techniques, the strongest causes of sea ice extent are **sst (lag 0)** and **snowfall (lag 0)**.

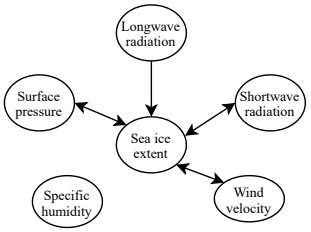


Fig. 11. The causal graph from physical knowledge [11].

c. Evaluation of Causal Graphs

The causal discovery approaches are evaluated by TPR, FPR, and precision which are introduced in Section IV-C. Table III shows the evaluation scores. Since we are interested in causal relationships between sea ice extent and other variables, we only look into edges related to sea ice extent, where sea ice extent is either the cause or the effect. Also, since there are only six variables are used in [11], the causal discovery results are evaluated based on variables including {*longwave radiation, shortwave radiation, wind velocity, surface pressure, specific humidity, sea ice extent*}. The physical knowledge-based causal graph is shown as Figure 11. From the table, we can see that, GES generates the best performance among IID methods with the highest precision and lowest FPR; while among time-dependent methods, VarLinGAM performs better than other methods considering the precision score.

TABLE III
EVALUATION ON CAUSAL DISCOVERY RESULTS BY PRECISION, TPR, AND FPR.

	IID methods			Time-Dependent methods		
	PC	GES	FCI	VarLiNGAM	PCMCI	PCMCI+
Precision	0.67	**1.00**	0.80	**0.75**	0.55	0.67
TPR	0.22	0.33	**0.36**	0.67	0.67	0.22
FPR	0.20	**0.00**	0.33	0.40	1.00	**0.20**

From these generated causal graphs, we identified the edges to analyze causal inference. Table IV and Table V show the two most possible causes of sea ice extent found by IID and time-dependent methods, respectively. We used the dimensional causality (DC) method to validate the causal edges between the $sst \rightarrow sea\_ice\_extent$ and $snowfall \rightarrow sea\_ice\_extent$. To compute the result of the DC method for the $sst \rightarrow sea\_ice\_extent$ we used embedded dimension 9 and τ (time delay) 1. For $snowfall \rightarrow sea\_ice\_extent$, the value of the embedded dimension is 9 and the τ is 0. Figure 12 shows that the DC method generated the perfect probability for the causal edge from sst to sea_ice_extent, but the probability assigned for the causal edge from $snowfall$ to sea_ice_extent is not aligned with causal knowledge. Moreover, the DC method generated a bidirectional causal link between the $snowfall$ and the sea_ice_extent with a higher probability. In the next section, we will analyze causal inference comparing the strength of causal effects for identified causes to sea ice extent.

TABLE IV
CAUSAL EDGES FOUND BY IID METHODS PC, GES, AND FCI.

Cause	Covariates	Effect
Sea surface temperature	Specific humidity, Air temperature	Sea ice extent
Air temperature	Sea surface temperature, Specific humidity	Sea ice extent

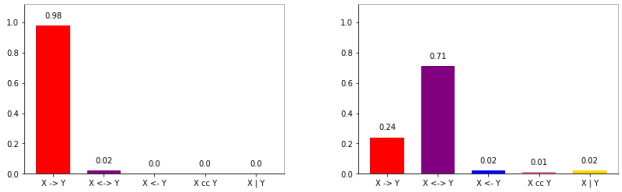
B. Causal Inference Results

a. Time-Independent Causal Inference

For time-independent causal inference, we use an adjustment-based causal inference to estimate the effects of

TABLE V
CAUSAL EDGES IDENTIFIED BY TIME-DEPENDENT METHODS PCMCI, PCMCI+, AND VARLINGAM.

Cause	Time lag	Covariates	Effect
Sea surface temperature	0	Specific humidity, Surface pressure	Sea ice extent
Snowfall	0	Rainfall, Longwave radiation	Sea ice extent



(a) *sst* ⟶ *sea_ice_extent* (b) *snowfall* ⟶ *sea_ice_extent*

Fig. 12. Validation of causal links using dimensional causality (DC) method.

sea surface temperature and air temperature with adjusted variables shown in Table IV, and a linear regression model is used to predict sea ice extent. The causal effect of sea surface temperature on sea ice extent is 72335, and the causal effect of air temperature on sea ice extent is 65486, which shows that the causal effect of sea surface temperature is stronger than air temperature.

b. Time-Dependent Causal Inference

We concluded from causal discovery results that sea ice extent is influenced by sea surface temperature (sst) and snowfall values. To measure the strength of the causal effect of these two causal variables on sea ice extent, we applied various state-of-the-art causal inference methods, including conditional instrumental variable, variable-lag granger causality, etc. To estimate the causal effect of sea surface temperature on sea ice extent using the conditional instrumental variable (CIV) method, $CIB_{sst \rightarrow sea\_ice\_extent}(I|B)$, we have considered $I = \{specific\_humidity, surface\_pressure\}$ and $B = \{rainfall\}$ and the β coefficient estimated by this method is 72616.73. The causal effect of snowfall to sea ice extent $CIB_{snowfall \rightarrow sea\_ice\_extent}(I|B)$ is measured using $I = \{rainfall, longwave\_radiation\}$ and $B = \{surface\_pressure, shortwave\_radiation\}$ and the estimated β coefficient is 13857571.99. Using these β coefficients and the value of the cause variable we can estimate the value of sea ice extent for any given instance of the time series with some fraction of error.

The results generated by the Variable-lag Granger causality and Transfer Entropy method are depicted in Table VI. Here all three methods have similar results for the causal edge $sst \rightarrow sea\_ice\_extent$ and using the method of BIC score calculation we can estimate the values of sea_ice_extent, using the value of sst. On the other hand, the Granger causality method failed to quantify the causal effect from $snowfall$ to sea_ice_extent, but both the VL-Granger causality and VL-Transfer Entropy methods detected the causal effect successfully.

TABLE VI
CAUSAL EFFECT FOUND BY VARIABLE-LAG (VL) GRANGER CAUSALITY AND TRANSFER ENTROPY METHOD.

Methods	$sst \rightarrow sea\_ice\_extent$	$snowfall \rightarrow sea\_ice\_extent$
Granger	X causes Y?=1 (BICDiffRatio = 0.529)	X causes Y?=0 (BICDiffRatio = 0.319)
VL-Granger	X causes Y?=1 (BICDiffRatio = 0.529)	X causes Y?=1 (BICDiffRatio = 0.812)
VL-Transfer Entropy	X causes Y?=1 (TEratio = 10.036)	X causes Y?=1 (TEratio = 2.003)

VI. DISCUSSION AND CHALLENGES

There is a common scenario for most of the real-world dynamical systems that we can monitor the system using different sensors and collect data from the system, but there is no structural equation to express this data. Also, it is not possible to turn off any physical force/part of the system and measure other parameters without the influence of that particular intervened component. Therefore, there is no mechanism to generate factual and counterfactual data from a dynamic system. The Arctic Sea Ice is a real-world dynamic system, so we have tried to analyze the causal effect between the various components of the system using a diverse set of causality methods. Some of these methods, such as Dragonnet, CEVAE, T-learners, and Time Series Deconfounder, require factual and counterfactual data for treatment variables in order to measure the causal effect of these variables on target variables. We did not apply these methods to our dataset due to the limitations of the counterfactual data. We are currently investigating possible strategies for generating counterfactual data for all treatment variables in the Arctic Sea Ice system.

VII. CONCLUSIONS

The causal discovery and causal inference methods are applied to the Arctic Sea Ice dataset to find the causes of sea ice extent and estimate the causal strength. The results of causal discovery methods show differences in inferred causes, and at the same time show the challenge of time-varying causal discovery. Though different causal discovery models generated different causal directions, the GES method works better for our dataset from applied IID methods and on the other hand, the VarLiNGAM method from time-dependent category generated a better temporal causal graph. From the results of causal inference models, we found that the conditional instrumental variable (CIV) method provided an estimate of the causal effect coefficient β between cause and effect in the dataset. In addition, the Variable-lag Granger causality method can determine the causal inference strength that is not determined by the regular Granger method. The limitations of this study are: (1) For causal inference, we need to find a more data-driven approach to decide the covariates, such as propensity score. (2) We need to find suitable evaluation metrics to compare results generated from causal inference methods.

For future work, besides solving the above limitations and problems, we plan to study the methods that we did not successfully use, such as deep learning-based causal inference methods. Also, we consider how the conclusions can help machine learning or deep learning-based study on the Arctic Sea Ice.

REFERENCES

[1] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–37, 2020.

[2] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search*. MIT press, 2000.

[3] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[4] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: journal of the Econometric Society*, pp. 424–438, 1969.

[5] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer, "Estimation of a structural vector autoregression model using non-gaussianity." *Journal of Machine Learning Research*, vol. 11, no. 5, 2010.

[6] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science advances*, vol. 5, no. 11, p. eaau4996, 2019.

[7] J. Runge, "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets," in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 1388–1397.

[8] M. Nauta, D. Bucur, and C. Seifert, "Causal discovery with attention-based convolutional neural networks," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, p. 19, 2019.

[9] B. Bussmann, J. Nys, and S. Latré, "Neural additive vector autoregression models for causal discovery in time series," in *International Conference on Discovery Science*. Springer, 2021, pp. 446–460.

[10] O. Faruque, S. Ali, X. Zheng, and J. Wang, "Ts-causalnn: Learning temporal causal relations from non-linear non-stationary time series data," 2024. [Online]. Available: https://arxiv.org/abs/2404.01466

[11] Y. Huang, M. Kleindessner, A. Munishkin, D. Varshney, P. Guo, and J. Wang, "Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere," *Frontiers in big Data*, vol. 4, 2021.

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[15] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3076–3085.

[16] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," *Advances in neural information processing systems*, vol. 30, 2017.

[17] J. Kerman, P. Wang, and J. Vaver, "Estimating ad effectiveness using geo experiments in a time-based regression framework," 2017.

[18] F. Menchetti, F. Cipollini, and F. Mealli, "Estimating the causal effect of an intervention in a time series setting: the c-arima approach," *arXiv preprint arXiv:2103.06740*, 2021.

[19] S. Ali, O. Faruque, Y. Huang, M. O. Gani, A. Subramanian, N.-J. Schlegel, and J. Wang, "Quantifying causes of arctic amplification via deep learning based time-series causal inference," in *2023 International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2023, pp. 689–696.

[20] I. Bica, A. M. Alaa, J. Jordon, and M. van der Schaar, "Estimating counterfactual treatment outcomes over time through adversarially balanced representations," *arXiv preprint arXiv:2002.04083*, 2020.

[21] A. Shojaie and E. B. Fox, "Granger causality: A review and recent advances. arxiv 2021," *arXiv preprint arXiv:2105.02675*, 2021.

[22] I. Ebert-Uphoff and Y. Deng, "Causal discovery for climate research using graphical models," *Journal of Climate*, vol. 25, no. 17, pp. 5648–5665, 2012.

[23] J. J. Nichol, M. G. Peterson, G. M. Fricke, and K. J. Peterson, "Learning why: Data-driven causal evaluations of climate models."

[24] C. Deser, R. Tomas, M. Alexander, and D. Lawrence, "The seasonal atmospheric response to projected arctic sea ice loss in the late twenty-first century," *Journal of Climate*, vol. 23, no. 2, pp. 333–351, 2010.

[25] C. Raible and R. Blender, "Northern hemisphere midlatitude cyclone variability in gcm simulations with different ocean representations," *Climate Dynamics*, vol. 22, no. 2, pp. 239–248, 2004.

[26] D. Handorf, R. Jaiser, K. Dethloff, A. Rinke, and J. Cohen, "Impacts of arctic sea ice and continental snow cover changes on atmospheric winter teleconnections," *Geophysical Research Letters*, vol. 42, no. 7, pp. 2367–2377, 2015.

[27] M. Kretschmer, D. Coumou, J. F. Donges, and J. Runge, "Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation," *Journal of climate*, vol. 29, no. 11, pp. 4069–4081, 2016.

[28] S. M. Samarasinghe, "Causal inference using observational data-case studies in climate science," Ph.D. dissertation, Colorado State University, 2020.

[29] D. D. Zhang, H. F. Lee, C. Wang, B. Li, Q. Pei, J. Zhang, and Y. An, "The causality analysis of climate change and large-scale human crisis," *Proceedings of the National Academy of Sciences*, vol. 108, no. 42, pp. 17 296–17 301, 2011.

[30] P. Nowack, J. Runge, V. Eyring, and J. Haigh, "Causal networks for climate model evaluation and constrained projections, nat. commun., 11, 1415," 2020.

[31] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí *et al.*, "Inferring causation from time series in earth system sciences," *Nature communications*, vol. 10, no. 1, pp. 1–13, 2019.

[32] C. G. Peter Spirtes and R. Scheines, "Causation, prediction, and search, 2nd edn. cambridge," *MA:MIT press*, 2001.

[33] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of Machine Learning Research, 507–554. doi: 10.1162/153244303321897717*, 2003.

[34] P. L. Spirtes, C. Meek, and T. S. Richardson, "Causal inference in the presence of latent variables and selection bias," *arXiv preprint arXiv:1302.4983*, 2013.

[35] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer, "Estimation of a structural vector autoregression model using non-gaussianity," *Journal of Machine Learning Research*, vol. 11, no. 56, pp. 1709–1731, 2010. [Online]. Available: http://jmlr.org/papers/v11/hyvarinen10a.html

[36] F. Takens, "Detecting strange attractors in turbulence." Springer Berlin, Heidelberg, 1981.

[37] Z. Benko, A. Zlatniczki, D. Fabó, A. Sólyom, L. Oss, A. Telcs, and Z. Somogyvári, "Exact inference of causal relations in dynamical systems," 08 2018. [Online]. Available: https://arxiv.org/pdf/1808.10806v2.pdf

[38] J. M. D. S. R. Z. Christos Louizos, Uri Shalit and M. Welling, "Causal effect inference with deep latent-variable models," *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[39] C. Shi, D. Blei, and V. Veitch, "Adapting neural networks for the estimation of treatment effects," *Advances in neural information processing systems*, vol. 32, 2019.

[40] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the national academy of sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.

[41] C. Amornbunchornvej, E. Zheleva, and T. Berger-Wolf, "Variable-lag granger causality and transfer entropy for time series analysis," 02 2020.

[42] I. Bica, A. Alaa, and M. Schaar, "Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders," 02 2019.

[43] N. Thams, R. Søndergaard, S. Weichwald, and J. Peters, "Identifying causal effects using instrumental time series: Nuisance iv and correcting for the past," *arXiv preprint arXiv:2203.06056*, 2022.

[44] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine learning*, vol. 65, no. 1, pp. 31–78, 2006.