TS-CausalNN: Learning Temporal Causal Relations from Non-linear Non-stationary Time Series Data

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

Sahara Ali
Department of Data Science
University of North Texas
Texas, USA
Sahara.Ali@unt.edu

Xue Zheng
Climate Science Section
Lawrence Livermore National Laboratory
California, USA
zheng7@llnl.gov

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

***Abstract*—The growing availability and importance of time series data across various domains, including environmental science, epidemiology, and economics, has led to an increasing need for time-series causal discovery methods that can identify the intricate relationships in the non-stationary, non-linear, and often noisy real world data. However, the majority of current time series causal discovery methods assume stationarity and linear relations in data, making them infeasible for the task. Further, the recent deep learning-based methods rely on the traditional causal structure learning approaches making them computationally expensive. We propose a Time-Series Causal Neural Network (TS-CausalNN) - a deep learning technique to discover contemporaneous and lagged causal relations simultaneously. Our proposed architecture comprises (i) convolutional blocks comprising parallel custom causal layers, (ii) acyclicity constraint, and (iii) optimization techniques using the augmented Lagrangian approach. The proposed model learns non-stationary features from input data through learning and aggregating nonlinear functions. Using these learned nonlinear features, the parallel custom causal convolution 2D layers learn causal connections between different variables. Through experiments on multiple synthetic and real world datasets, we demonstrate the effectiveness of our proposed approach as compared to several state-of-the-art methods. The inferred graphs for the real world dataset are in good agreement with the domain understanding.**

***Index Terms*—Causal Discovery, Time Series Data, Non-Stationarity, Non-Linear, Neural Network**

I. Introduction

Multivariate time series data generated by different natural systems, such as climate and environment, can possess different features like non-linearity, non-stationarity, presence of different noise categories, and autocorrelation [4]. These intricate characteristics of time series data ingraft complex challenges in understanding the dependencies of different components of these natural systems. One popular approach to simplify the understanding of large multivariate time series datasets is to graphically represent the data generation model using directed acyclic graphs (DAGs), which is a very convenient way to express complex systems in a highly interpretable manner and also provide causal insights into the underlying processes [3]. DAG representation of a system plays a vital role in decision-making and future condition prediction in different applications like causal inference [2], [5], neuroscience [1], medicine [28], economics [26], finance [25], and machine learning [27]. Learning DAG through causal discovery from observational time series data is very challenging when controlled experiments with different population sub-groups are not possible or unethical [5], [16].

Several state-of-the-art methods have been developed for causal discovery from temporal data based on constraint-based and score-based methodologies. Constraint-based causal discovery methods [13], [14], [18]–[20] learn the conditional independencies of the data using different tests and construct the DAG to reflect these independencies. One major limitation of the conditional independence test is that it requires a large number of samples to generate reliable test scores [17]. Score-based causal discovery methods use a score function to quantify the predicted causal graph based on the adjacency matrix and try to optimize the score function by enforcing different graph constraints. The large search domain of the score function based on the adjacency matrix makes the optimization process very challenging and sometimes requires additional knowledge of the DAG. The combinatorial characteristic of score optimization was transformed into a continuous optimization problem by Zheng et al. [10], formulating an equivalent acyclicity constraint using the trace exponential of the predicted adjacency matrix. Through this continuous formulation of the acyclicity constraint, now causal graphs can be optimized using SOTA gradient-based optimizers. This drastically changes the pace of multivariate time series causal discovery using neural network-based approaches, and several methods have been proposed in recent years.

Time series causal discovery leveraging the power of neural networks has gained much attention as an active research area in several fields. However, the majority of existing methodologies are designed under the assumption of stationarity, a condition often violated in real world scenarios where

dynamic systems evolve periodically [4], [43]. Some existing approaches also require prior knowledge of the domain, like linearity, noise distribution, and parametric information. Motivated by the tremendous success of neural networks [34], in this paper, we propose Time-Series Causal Neural Network (TS-CausalNN) - a novel causal discovery method for temporal data using a convolutional neural network. The causal relationship between each child and its temporal parents is learned using a custom 2D convolution layer. Our proposed model is capable of capturing the causal structure from multivariate temporal data without any noise and data distribution assumptions. This method is designed to be applicable across diverse domains for multivariate time series data. The contributions of this paper are three-fold. First, we propose a 2D convolutional neural network layer to learn the causal relationships from the multivariate time series dataset. By utilizing the power of CNN, our proposed model can handle both the stationary and non-stationary data for linear and non-linear structural models with the presence of different noise distributions. The proposed parallel network architecture has not been used earlier. Second, the simplified optimization routine of the proposed model can identify lagged and contemporaneous causal links simultaneously. The integration of the acyclicity constraint and sparsity penalty into the optimization process helps to learn better causal graphs. Finally, we conduct extensive evaluations of the proposed model with state-of-the-art methods using synthetic and real world datasets. The proposed model achieves better evaluation scores for generated causal graphs for most of the cases, making TS-CausalNN a strong contender for time-series causal discovery.

II. Preliminaries

Temporal causal discovery learns directed acyclic graphs (DAG) from time series data. Each directed edge of the DAG represents the influence of the cause variable on the target variable. As the order of data is strictly maintained in the time series case, influences from cause variables to the target variable can only come from the same and previous timesteps, also called temporal precedence (the right side of Figure 1). This feature makes causal graph learning from time series data more challenging compared to independent and identically distributed (IID) data.

Let's consider a multivariate time series dataset $X = \{x^1, x^2, x^3, \ldots, x^n\}$ consisting of n variables, and each variable is measured for T timesteps. Variable x^i at a specific time point t can be caused by other variables at the same time point (t) and all variables from previous timesteps (0 to $t-1$). The effect from previous timesteps, also called lagged effects, can propagate from infinitely earlier time points, but for DAG learning purposes, we will consider a maximum time lag, l_{max}. So the set of possible cause variables of each time series x^i at time t is $PA_{x^i} \in [\{X_{(t-l_{max})}, X_{(t-l_{max}-1)}, ..., X_{(t-1)}, X_t\} - x^i]$. The goal is to learn a causal graph $G(V, E)$ from the provided dataset such that its vertices resemble time-lagged and current time variables in the time series and its directed edges express parent-to-child causal links. So the vertices and edges are denoted as $V = \{X_{(t-l_{max})}, X_{(t-l_{max}-1)}, ..., X_{(t-1)}, X_t\}$, $E = \{(V_i, V_j) : V_i, V_j \in \{X_{(t-l_{max})}, X_{(t-l_{max}-1)}, ..., X_{(t-1)}, X_t\}\}$, respectively. Let the weighted adjacency matrix of causal graph G be denoted by $W \in R^{(n\times(l_{max}+1))\times n}$, which contains both the time-lagged and instantaneous part of the causal links. The structural equation model of the time series can be defined as:

$$X_t = f_W(X_{(t-l_{max})}, X_{(t-l_{max}-1)}, ..., X_{(t-1)}, X_t) + e_t, \quad (1)$$

where the noise term e_t can be of any type, independent of cause and effect, and the structural function $f_W()$ can be any linear or non-linear data generation process. The learning of the target causal structure from the input time series depends on the following assumptions.

Assumption 1 (Markov and Faithfulness): Assume $X_i, i \in \{1, ..., n\}$ is Markov and faithful to the true/generated causal graph G [43].

Assumption 2 (Causal Sufficiency): We assume that there are no unobserved confounders in the data generation process.

Assumption 3 (Causal Consistency): We assume that time-lagged and instantaneous causal relations between the variables are consistent through all time steps.

Assumption 4 (Acyclicity): This assumption states that there are no causal paths that begin and end at the same node.

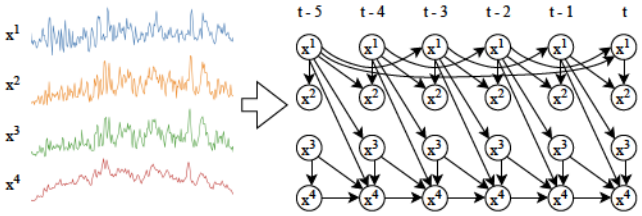


Fig. 1. Temporal causal graph learned (right) from multivariate time series data (left) with causal links from the same and previous timesteps. Each node in the graph represents one variable at a specific timestep (t is the current time and $t-i$ means previous timesteps). A directed edge denotes a causal relationship between cause and effect.

In the time-lagged part of the adjacency matrix W, the nodes from previous l_{max} timesteps will always be the source/cause node of the lagged causal links. So, we do not need to explicitly focus on the acyclicity of the time-lagged part of W. However, for the contemporary part of the W at t, each node can serve as both the source and target of causal links, where we have to maintain the acyclicity of the DAG. Several methods can be used to recover the adjacency matrix W of the causal graph from time series data, like search-based, constraint-based, score-based, graph-based methods, etc. To formulate this combinatorial search problem into a continuous optimization nature, Zheng et al. [10] proposed an algebraic representation of the acyclicity constraint with the matrix exponential. This new representation of acyclicity opens a window to learn the DAG using generic continuous optimization routines like neural networks by minimizing the score. However, simultaneously learning the lagged

and contemporaneous parts of the adjacency matrix is very challenging for complex datasets. Since any variable might be the cause of another effect variable, cycles can occur in the contemporaneous part of the adjacency matrix. Consequently, these two parts of the adjacency matrix require a different set of optimization criteria.

III. Proposed Methodology

We can consider the causal graph generation task as an unsupervised learning process of the adjacency matrix W given the multivariate time series data $X = \{x^1, x^2, x^3, \ldots, x^n\}$ of T observations. To learn the directed adjacency matrix W of the temporal causal graph G, we propose a neural network-based unsupervised model. The proposed time series causal neural network model will learn the instantaneous $(X_t \rightarrow X_t)$ and time-lagged $(\{X_{(t-l_{max})}, X_{(t-l_{max}-1)}, ..., X_{(t-1)}\} \rightarrow X_t)$ causal links of W for maximum time lag $(l_{max} > 0)$ in the same gradient propagation using a single neural network. To perform the temporal causal learning, we propose a custom 2D causal convolution layer.

A. Time Series Causal Neural Network

As illustrated in Figure 2, our proposed TS-CausalNN model consists of two separate blocks of 2-dimensional convolution layers, which helps its simplicity and efficiency. The first 2D convolution layer of the model transforms input data into a latent representation, which helps the model to learn non-linear relationships present in the data generation process. The Causal Conv2D block contains n parallel custom causal layers to learn the time-lagged and instantaneous causal relationships of each input variable to its parent variables. Following a similar analogy used by Zheng et al. [11], DAG-GNN [12] and NOTEARS-MLP [10], we will learn the causal links from the parameters of the Conv2D block. The links learned by this causal block are always unidirectional and the parallel blocks help to learn the causal links for each input variable independently. Finally, the results of each parallel causal layer are aggregated to generate the model output, which will be used in the optimization process of the learned causal graph.

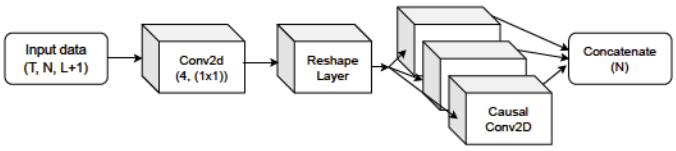


Fig. 2. Proposed TS-CausalNN model architecture to learn full temporal causal graph. The graph is learned from the parallel Causal Conv2d layers.

Our proposed Causal Convolution 2D (Causal Conv2D) layer takes input in a structure similar to that shown in the full causal graph on the right side of Figure 1, with lagged data followed by the current time point data. Each Causal Conv2D layer is designed to learn the causal links of an input variable for example x^1 from all possible parents $PA_{x^1} \in \{x^1_{(t-l_{max})}, x^1_{(t-l_{max}-1)}, ..., x^1_{(t-1)}, x^2_{(t-l_{max})}, x^2_{(t-l_{max}-1)}, ..., x^2_{(t-1)}, x^2_t, ..., x^n_{(t-l_{max})}, x^n_{(t-l_{max}-1)}, ..., x^n_{(t-1)}, x^n_t\}$ of that variable. The variable itself cannot be included in the set of its parent variables. Let's assume we have a time series dataset with 4 variables $X = \{x^1, x^2, x^3, x^4\}$, and for lagged effects consider the maximum time lag $l_{max} = 4$. So the input data will be a matrix of size (4×5), one row for each variable and $l_{max} + 1 = 5$ column for lagged and contemporaneous variables. To learn the temporal causal graph for these 4 variables, as shown in Figure 3, we have to employ 4 parallel Causal Conv2D layers, one layer for each variable. Each of these layers predicts the expectation of the target variable at timestep t given all lagged and instantaneous parents (Equation 2). To exclude the target variable from the corresponding list of parents, its weight is set to zero.

$$E[x^i | PA_{x^i}] = f_{W^{x^i}}(PA_{x^i}) \tag{2}$$

Here $f_{W^{x^i}}()$ is the function learned in the Causal Conv2D layer and W^{x^i} is the weight parameters of that layer. Motivated by NOTEARS-MLP [10] and DAG-GNN [12], we derive the adjacency matrix of the causal DAG from weight parameters of the Causal Conv2d layers. The weight parameter of the layer for a target variable represents the strength of the causal link from its parent. The weight parameter $W^{x^k}_{ij} = 0$ means the target variable x^k is independent of the cause variable x^i at timestep j. If $W^{x^k}_{ij} > 0$, then the variable x^k has a causal edge from the parent variable x^i at time lag j. After learning the final weight parameters of the target variable, we apply a thresholding operation to prune edges with weak dependency strength, $W^{x^k}_{\omega} = (W^{x^k} > \omega)$, where ω is the threshold value. Finally, the weight parameters of all variables after thresholding are concatenated to generate the adjacency matrix of the final causal graph.

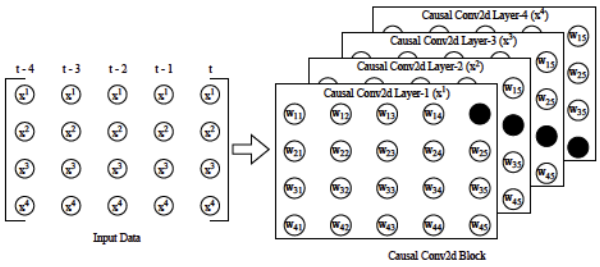


Fig. 3. Proposed custom Causal Conv2D Layers.

B. Acyclicity Constraint

Maintaining the acyclic property of the adjacency matrix makes it challenging to learn causal graphs from time series data. Using a continuous optimization process like a neural network to learn the causal graph does not guarantee the acyclicity of the learned adjacency matrix. To impose the acyclicity restriction in the adjacency matrix of the learned causal DAG, we will use a similar equality constraint $h(W) = 0$ as Zheng et al. [10]. The function $h(W)$ is defined using the trace exponential $(tr())$ of the elementwise product of the adjacency matrix with itself, $h(W) = tr(e^{W \circ W}) - n$. Here n

is the number of variables in the dataset. We cannot use the learned adjacency matrix W of the proposed TS-CausalNN method directly in this equality function because W contains both the time-lagged $(t - l_{max}, t - l_{max} - 1, \ldots, t - 1)$ and contemporaneous (t) edges of the causal graph. The time-lagged edges of the causal graph will always go forward in time, from previous timesteps to current timestep t; therefore, we do not need to compute acyclicity for this part of the graph. We have to apply the acyclicity constraint on the contemporaneous part of the adjacency matrix, W^t. Hence, the function can be stated below, and the equality will be satisfied if and only if the contemporaneous part of the adjacency matrix (W^t) is acyclic.

$$h(W^t) = tr(e^{W^t \circ W^t}) - n = 0 \tag{3}$$

C. Optimization

The goal of the proposed causal discovery method is to learn the adjacency matrix W of the DAG from the given time-series dataset. The adjacency matrix contains edges from the time-lagged variables and the contemporaneous variables to the target variables. So, the proposed TS-CausalNN model estimates the target variables using all possible parents and finds the causal influence (W) of each parent on the target, where the contemporaneous part of W^t must satisfy the acyclicity. The optimization objective of the model is to minimize the least square loss (L) of the target variables with the acyclicity constraint on W^t.

$$\min_W L(W) \text{ subject to } h(W^t) = 0 \text{ for acyclicity,} \tag{4}$$

$$\text{where } L(W) = \frac{1}{T}||X - WX||_F^2, \tag{5}$$

$$\text{and } h(W^t) = tr(e^{W^t \circ W^t}) - n = 0 \tag{6}$$

The function $h(W^t)$ will be equal to 0 if and only if the corresponding sub-graph of matrix W^t does not have any cycle. We cannot directly integrate the equality constraint of the acyclicity into the continuous optimization function. But the equality constraint $h(W^t) = 0$ can be solved using continuous optimization after converting this into an unconstrained problem [10]. Therefore, we use the augmented Lagrangian method to solve the equality constraint problem. An additional penalty is incorporated with the objective function to enforce the sparsity of the adjacency matrix using the $L1$ norm of W. The least square loss and the acyclicity constraint (Equations 5 and 6) will try to increase the weight parameter to minimize the loss. On the contrary, the L1 norm of W will try to reduce weight values to zero to keep a minimum number of non-zero entries. By working contrarily to each other, the loss function will be optimized in an equilibrium way. Hence, the final unconstrained objective function is:

$$\min_W \left[L(W) + \frac{\rho}{2}|h(W^t)|^2 + \alpha h(W^t) + \lambda||W||_1 \right], \tag{7}$$

where the 2nd and 3rd terms are the augmented Lagrangian for the acyclicity constraint, α is the Lagrange multiplier, $\rho > 0$ is the penalty parameter of the augmented Lagrangian, and λ is the sparsity penalty parameter. This objective function can be minimized using any state-of-the-art continuous optimizer. An excellent characteristic of the augmented Lagrangian approach is its ability to accurately approximate the solution of a constrained problem using the solution of unconstrained problems by gradually increasing the penalty parameter ρ but not to infinity. Hence, we gradually increase ρ to minimize the unconstrained augmented Lagrangian, and the value of λ also has to be updated accordingly. The rule for updating ρ and λ based on the value of $h(W^i)$ is given by:

$$\rho^{i+1} = \begin{cases} (1+\beta)\rho^i, & \text{if } h(W^i) > \gamma h(W^{i-1}) \\ \rho^i, & \text{otherwise} \end{cases}, \tag{8}$$

$$\text{and } \alpha^{i+1} = \alpha^i + \rho^i h(W^i), \tag{9}$$

where $\beta > 0$ and $\gamma < 1$ are hyperparameters, and we find that $(\beta = 0.1, \gamma = 0.25)$ work better. The overall process of learning causal DAG is outlined in Algorithm 1.

Algorithm 1: TS-CausalNN Algorithm

Input: Multivariate Time Series Data
$X = \{x^1, x^2, x^3, ..., x^n\}$
Output: Adjacency Matrix of the causal graph W

```
Instantiate and compile the Causal Neural Network model
for each iteration do
    Train the Neural Network model
    Compare the acyclicity loss with the previous iteration
    Update the penalty coefficient of the Lagrangian method (ρ, α)
Return the adjacency matrix W
```

IV. Experimental Setup

We mention the dataset and evaluation criterion used for performance comparison in this section. Our model is developed using TensorFlow Keras, and all experiments are conducted on Google Colab Runtime with CPU for easy reproducibility. Fixed seed values are used for each experiment to make the experimental results reproducible.

A. Synthetic Datasets

To evaluate the performance of our proposed causal discovery method we have used synthetic datasets. As we know the ground truth causal graph for the synthetic datasets, we can measure and compare the learned causal graph easily. We generate a time series dataset consisting of four non-linear variables using Gaussian white noise ε (Dataset-1). The mathematical description of each variable is given in Equations 10 to 13. The non-linear characteristic is incorporated in the generation of synthetic dataset to mimic the dynamic properties of real world natural system data. The corresponding true causal graph for the time-series data is given in Figure 4a.

$$S1_t = 2\{cos(\frac{t}{10}) + log(|S1_{t-2} - S1_{t-5}| + 1)\} + 0.1\varepsilon 1 \tag{10}$$

$$S2_t = 12e^{\frac{S1_{t-1}^2}{2}} - 4e^{\frac{S1_t^2}{2}} + \varepsilon 2 \tag{11}$$

$$S3_t = -10.5e^{\frac{-S1_{t-1}^2}{2}} + \varepsilon 3 \tag{12}$$

$$S4_t = -11.5e^{\frac{-S1_{t-1}^2}{2}} + 13.5e^{\frac{-S3_{t-1}^2}{2}} + 1.2e^{\frac{-S4_{t-1}^2}{2}} - 5e^{\frac{-S3_t^2}{2}} + \varepsilon 4 \tag{13}$$

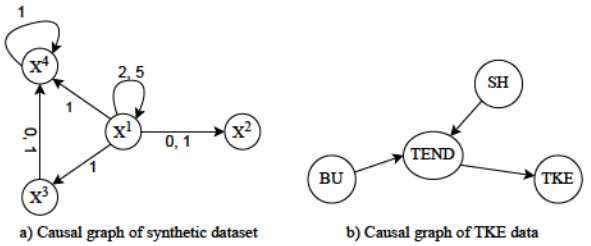


Fig. 4. Causal graph of (a) our synthetic datasets and (b) the real world Turbulence Kinetic Energy (TKE) dataset.

We have also generated another synthetic dataset (Dataset-2) from the causal graph provided in Figure 4a without the cosine part of the data generation formulas. For this case, we used the Poisson distribution to incorporate random noise into the dataset. The goal is to find the performance and robustness of causal discovery methods for the time series dataset without having wave-shape-like features with non-Gaussian noise. The formula used to generate this dataset is given below.

$$S1_t = 0.7e^{\frac{-S1_{t-2}^2 \times S1_{t-5}^2}{2}} + \varepsilon 1 \tag{14}$$

$$S2_t = 2e^{\frac{S1_{t-1}^2}{2}} + 0.5e^{\frac{S1_t^2}{2}} + \varepsilon 2 \tag{15}$$

$$S3_t = -5.05e^{\frac{-S1_{t-1}^2}{2}} + \varepsilon 3 \tag{16}$$

$$S4_t = -1.15e^{\frac{-S1_{t-1}^2}{2}} + 2.35e^{\frac{-S3_{t-1}^2}{2}} + 1.5e^{\frac{-S4_{t-1}^2}{2}} + 3e^{\frac{-S3_t^2}{2}} + \varepsilon 4 \tag{17}$$

Before applying the proposed causal discovery method, we performed some data preprocessing to ensure the quality and consistency of the input data. Different variables of the dataset have different scales of values. All time series data were normalized within the scale of 0 to 1 to mitigate the impact of scale differences. Then l_{max} previous timesteps' data of each current timestep is concatenated in front of it to make the dataset size $(T, n, (l_{max} + 1))$. The preprocessed data were applied to the input convolution 2D layer of the proposed model.

B. Simulation and Real World Datasets

One real world simulation data, namely Turbulence Kinetic Energy (TKE), and another real observational Arctic Sea Ice data were used to evaluate our work. These natural datasets exhibit high variability, non-stationarity, and complex data interactions. Owing to these characteristics, we evaluate our model performance on these datasets to assess how well the proposed model performs on such complex data.

TKE refers to the mean kinetic energy per unit mass associated with eddies in turbulent flow [21]. It can be measured by the root-mean-square (RMS) velocity fluctuations (unit: m^2s^{-2}). The temporal TKE data used in this study represent the evolution during a typical cumulus-topped boundary layer day (local time 05:00 – 18:00) over the DOE Atmospheric Radiation Measurement (ARM) Southern Great Plains Central Facility. This data file is generated from an idealized numerical simulation using the Weather Research & Forecasting (WRF) Model [22] with modifications from the Large-Eddy Simulation (LES) Symbiotic Simulation and Observation (LASSO) activity, which is developed through the US Department of Energy's ARM facility [23], [24]. Besides TKE, the model also generated the budget terms determining the temporal change in TKE. The major budget terms include the TKE vertical shear production term (SH, m^2s^{-3}), the TKE buoyancy production term (BU), and the TKE turbulent and pressure transport term (TR). All these budget terms form the net temporal change term of TKE ($TEND, m^2s^{-3}$). If $TEND$ is positive (negative), TKE will increase (decrease) in the next timestep. Figure 4b illustrates how these terms relate to one another through a directed graph.

Arctic sea ice is one of the important components of the world's climate system that has a great impact on the increase of extreme weather events [38]–[42] and the ice is melting rapidly due to various atmospheric conditions [36], [37]. Huang et al. [35] conducted a causal discovery analysis to uncover the links between the arctic sea ice and the atmosphere. We use the same 11 atmospheric variables with the sea ice extent employed in [35] and obtained from the ERA-5 global reanalysis data product in monthly averages from 1980 to 2018 over the Arctic region of 60N. The causal relationship between these variables and the sea ice extent based on the physics and microphysics literature review discussed in [35].

C. Evaluation Metrics

To evaluate the performance of the time series causal discovery methods we use three standard evaluation metrics: Structural Hamming Distance (SHD), F1 Score, and False Discovery Rate (FDR). SHD of the directed graph represents the smallest number of edge corrections required to transform the predicted causal graph into the true causal graph. The edges can be corrected by adding new edges, removing existing edges, and changing the direction of an existing edge in the graph. A lower SHD represents better performance of the causal discovery method. The F1 Score calculates the harmonic mean of precision and recall, providing a balanced measure of performance. The F1 score ranges from 0 to 1 and a higher value means a better prediction of the true graph. FDR explains the rate of predicted wrong edges from all predicted edges considering the direction of each edge. This is the inverse measure of the precision of the prediction.

V. Results

In this section, we present the comparative results of the time series causal discovery between the proposed method

TABLE I
COMPARISON OF THE PREDICTED SUMMARY CAUSAL GRAPH BY DIFFERENT METHODS FOR SYNTHETIC DATASETS.

METHOD	DATASET-1			DATASET-2		
	SHD	F1	FDR	SHD	F1	FDR
PCMCI [13]	10	0.54	0.62	7	0.63	0.53
PCMCI+ [14]	10	0.54	0.62	6	**0.66**	0.50
NOTEARS-MLP [11]	6	0.25	0.50	**4**	0.50	**0.00**
NTS-NOTEARS [15]	**3**	**0.76**	**0.28**	5	0.44	0.33
DAG-GNN [12]	5	0.44	0.33	8	0.50	0.60
DYNOTEARS [8]	4	0.74	0.40	7	0.36	0.60
TCDF [48]	5	0.54	0.40	5	0.54	0.40
PROPOSED	**3**	**0.76**	**0.28**	**4**	0.60	0.25

TABLE II
COMPARISON OF THE PREDICTED TEMPORAL FULL CAUSAL GRAPH BY DIFFERENT METHODS FOR SYNTHETIC DATASETS.

METHOD	DATASET-1			DATASET-2		
	SHD	F1	FDR	SHD	F1	FDR
PCMCI [13]	69	0.16	0.90	22	0.38	0.74
PCMCI+ [14]	49	0.16	0.90	11	**0.62**	0.55
NOTEARS-MLP [11]	13	0.58	0.59	20	0.44	0.70
NTS-NOTEARS [15]	12	0.45	0.61	8	0.33	0.33
DAG-GNN [12]	12	0.00	1.00	18	0.28	0.80
DYNOTEARS [8]	13	0.51	0.61	18	0.25	0.80
TCDF [48]	10	0.28	0.60	8	0.42	0.40
PROPOSED	**8**	**0.60**	**0.45**	**7**	0.46	**0.25**

and state-of-the-art methods. Several typical temporal causal discovery challenges have been considered here, like time-lagged and contemporaneous causal relations, different noise distributions, nonlinearity of data, and autocorrelation.

A. Baselines

To compare the results of the proposed method with state-of-the-art methods, we considered PCMCI+ [14], DYNOTEARS [8], NTS-NOTEARS [15], TCDF [48] PCMCI [13], NOTEARS-MLP [11], and DAG-GNN [12]. The first five methods can directly learn causal graphs for time series data. Though the other two methods were proposed for non-temporal data we used these methods due to their popularity and widespread usage in different domains. We transformed the lagged and instantaneous data into a long sequence so that we could apply the transformed dataset to the non-temporal methods to find the lagged and current time causal relationships. For each SOTA method, we tuned hyperparameters to get the best evaluation scores.

B. Quantitative Results

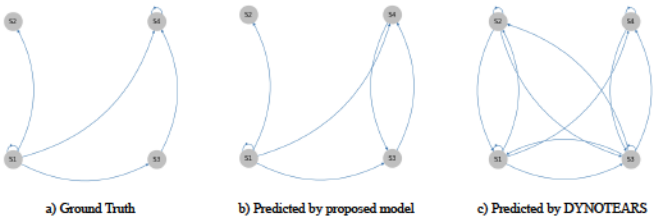
To evaluate the performance of the selected state-of-the-art methods we compared the predicted causal graph in both the summary and full temporal graph settings. The qualitative comparison of the baseline methods is reported in Tables I and II. The best results are marked in bold text and underlined values represent the second best score. To get better prediction results, we have applied each method multiple times on the same dataset and reported the best result. The comparative analysis of the summary graph (Table I) shows that the proposed method achieves the joint best scores for all three evaluation criteria for dataset-1. For synthetic dataset-2, NOTEARS-MLP and proposed method yields a lower score in terms of FDR, whereas the F1 score of the NOTEARS-MLP is less and FDR is zero. This means that the NOTEARS-MLP method generated less number of edges than the proposed method as a result missed some true edges in the predicted causal graph.

To compare the results of the predicted full causal graph, we considered both the directed edges and the time lag associated with each edge. From Table II, we can see that our proposed method got the best results for all three quality measures for dataset-1. For dataset-2, the proposed method yields the best F1 score and FDR, and the second-lowest SHD. For dataset-1, the TCDF method generated second best SHD, but the F1 score is quite low, means this method generated few causal edges and 60% of these edges are wrong. For dataset-2, the PCMCI+ method yields the best F1 score despite comparatively higher SHD and FDR than our proposed method. In this case, the PCMCI+ method generated more causal edges than proposed method, so it produced higher false causal edges, which leads to higher FDR. Observing the evaluation results of the full causal graph, we can find that the proposed model is capable of generating fewer false causal links compared to other baseline models.

C. Qualitative Results

To gain more insights into the discovered causal relationships, we conducted a qualitative analysis of the predicted causal graphs from synthetic dataset-1. Figure 5 illustrates the ground truth and predicted causal graphs of the proposed and DYNOTEARS methods using direct graphs. The nodes in the graph represent the variables in the dataset and each edge represents a causal relationship between the nodes. The self-loop in the causal graph represents that the previous timestep of the variable has a lagged effect on its present timestep. From the predicted causal graphs, we can see that the proposed model predicted the same number of edges as the ground truth graph but failed to predict edge $S4 \rightarrow S4$ and estimated a false edge $S3 \rightarrow S4$. On the other hand, the DYNOTEARS method failed to distinguish causal effects as mentioned in the ground truth graph. Instead, DYNOTEARS method predicts bidirectional effects between variables $(S1 \leftrightarrow S2)$,$(S1 \leftrightarrow S3)$,$(S2 \leftrightarrow S3)$, and $(S3 \leftrightarrow S4)$. Whereas $S1$ is the cause variable for the other three variables and there is no causal link between variable $S2$ and $S3$.

The full temporal causal graph of the data generation process of synthetic dataset-1 is visualized in the first plot of Figure 6. In this plot, columns represent the parent variables with a time lag and rows represent the child/target variables. The adjacency matrix of the true graph is sparse, which helps us to evaluate the performance of time series causal discovery methods for sparse causal graphs. From the predicted adjacency matrix of the proposed model, we can see that the causal



a) Ground Truth b) Predicted by proposed model c) Predicted by DYNOTEARS

Fig. 5. Comparison of ground truth summary causal graph with predicted graphs from DYNOTEARS and our proposed model for synthetic dataset-1.

links $\{S1(t-2) \rightarrow S1(t), S1(t-1) \rightarrow S3(t), S1(t-1) \rightarrow S2(t), S3(t-1) \rightarrow S4(t)\}$ are predicted correctly. However, the proposed model failed to predict the other three true causal links for dataset-1. On the other hand, the DYNOTEARS method predicted all edges from the true causal graph except $(S1(t-5) \rightarrow S1(t))$ and it predicted more false edges than the proposed model. The qualitative and quantitative analysis of the predicted results provides us with the intuition that most state-of-the-art models can work better on predicting dense causal graphs than sparse causal graphs.

a) Ground Truth

b) Predicted by proposed model

c) Predicted by DYNOTEARS model

Fig. 6. Visualization of ground truth full temporal causal graph of synthetic dataset-1 and the predicted graphs from DYNOTEARS and our proposed model. (The yellow cell means there is a directed edge from the column index to the row index.)

D. Results on Real World Data

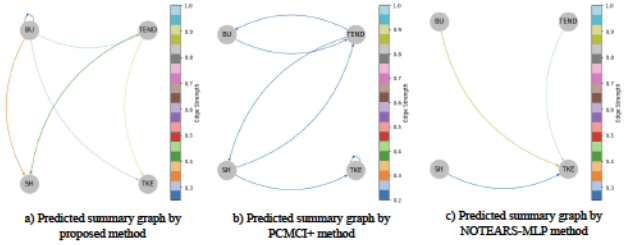
We applied the proposed and baseline models to the TKE and Arctic Sea Ice datasets to generate causal graphs with time lags of 5 and 12, referring to the very fast time evolution of planetary boundary layer turbulence in TKE data and annual seasonality in sea ice data, respectively. The evaluation results

TABLE III
COMPARISON OF THE PREDICTED SUMMARY CAUSAL GRAPH BY DIFFERENT METHODS FOR REAL WORLD DATASETS.

METHOD	TKE			ARCTIC SEA ICE		
	SHD	F1	FDR	SHD	F1	FDR
PCMCI [13]	9	0.18	0.87	62	0.31	0.68
PCMCI+ [14]	5	**0.44**	**0.66**	**50**	0.32	0.57
NOTEARS-MLP [11]	**4**	0.33	**0.66**	71	0.38	0.68
NTS-NOTEARS [15]	6	0.40	0.71	53	0.10	0.76
DAG-GNN [12]	7	0.22	0.83	66	0.21	0.76
DYNOTEARS [8]	8	0.00	1.00	65	0.21	0.75
TCDF [48]	6	0.25	0.80	51	0.21	0.63
PROPOSED	5	**0.44**	**0.66**	54	**0.49**	**0.56**

of the summary graphs generated by baseline models are summarised in Table III. For the TKE dataset, the proposed model performed similarly to PCMCI+ method. Although NOTEARS-MLP has a lower SHD score, its F1 score is lower than that of the proposed model. For the Arctic Sea Ice dataset, our proposed method achieves the best F1 and FDR scores. Although the SHD scores of a few baseline models were lower than those of the proposed model, their F1 and FDR values were not as good.

To compare generated causal graphs visually, we demonstrate the predicted causal summary graphs by PCMCI+, NOTEARS-MLP and the proposed method from the TKE dataset. We analyze the predicted causal graphs by comparing the domain graph represented in Figure 4. From the predicted graphs illustrated in Figure 7, we can see that the proposed method recovered the correct edges from $BU \rightarrow TEND$, $TEND \rightarrow TKE$, and the connection between each pair of nodes is unidirectional. On the other hand, the PCMCI+ method predicted bidirectional edges for node pairs $BU \rightarrow TEND$, and $SH \rightarrow TEND$, and missed the important edge from $TEND$ to TKE. However, NOTEARS-MLP predicted one correct edge $TEND \rightarrow TKE$ although it has a lower SHD score.



a) Predicted summary graph by proposed method b) Predicted summary graph by PCMCI+ method c) Predicted summary graph by NOTEARS-MLP method

Fig. 7. Visualization of predicted summary causal graphs of the TKE dataset using PCMCI+, NOTEARS-MLP, and our proposed model.

E. Ablation Study

To verify the effectiveness of the proposed model for temporal causal discovery, a comparative study of this model with its variant is shown here. The quantitative results of the generated summary causal graphs of these models are illustrated in Table V. For Causal Conv1D, we used a 1D

TABLE IV
COMPARISON OF THE PREDICTED SUMMARY CAUSAL GRAPH BY DIFFERENT METHODS FOR SYNTHETIC DATASETS WITH DIFFERENT SIGNAL-TO-NOISE RATIO (SNR).

METHOD	SYNTHTIC-1.1 SNR - 0.26			SYNTHTIC-1.2 SNR - 0.50			SYNTHTIC-1.3 SNR - 0.62		
	SHD	F1	FDR	SHD	F1	FDR	SHD	F1	FDR
PCMCI [13]	10	0.54	0.62	10	0.54	0.62	10	0.54	0.62
PCMCI+ [14]	8	**0.60**	0.57	10	0.54	0.62	9	0.54	0.62
NOTEARS-MLP [11]	6	0.25	0.50	8	0.20	0.75	6	0.25	0.50
NTS-NOTEARS [15]	NE	NE	NE	NE	NE	NE	6	0.25	0.50
DAG-GNN [12]	NE	NE	NE	NE	NE	NE	3	0.66	**0.00**
DYNOTEARS [8]	7	0.45	0.57	7	0.45	0.57	7	0.63	0.53
TCDF [48]	**5**	0.28	**0.00**	**3**	**0.66**	**0.00**	5	0.54	0.40
PROPOSED	**5**	0.54	0.40	4	**0.66**	0.33	**2**	**0.83**	0.16

TABLE V
PERFORMANCE COMPARISON BETWEEN CUSTOM CAUSAL CONV2D AND CONV1D MODELS.

DATASET	CAUSAL CONV1D			CAUSAL CONV2D		
	SHD	F1	FDR	SHD	F1	FDR
SYNTHETIC DATASE-1	7	0.46	0.57	3	0.77	0.28
SYNTHETIC DATASE-2	8	0.33	0.66	4	0.60	0.25
TKE	6	0.40	0.71	5	0.44	0.66
ARCTIC SEA ICE	63	0.42	0.63	54	0.49	0.56

variant of the proposed custom Causal Conv2D layer. To incorporate this Conv1D layer into the model, we flattened the latent representation of the earlier convolution layers of the model. The same training and optimization process was utilized for both models. The ablation study results show that the Custom Conv2D layer improves the causal graphs learning performance of the proposed model with a significant margin for each evaluation score. As we are studying multivariate time series data, 2D convolution layers can better learn the inherent features of the data generation than 1D convolution layers.

F. Robustness Analysis

To investigate the performance stability of the proposed model, we considered different features of datasets, including signal-to-noise ratio (SNR), noise distribution, non-stationary and stationary variable combinations. We generated three versions of the synthetic dataset-1 (Synthetic-1.1 with lower SNR, Synthetic-1.2 with medium SNR, and Synthetic-1.3 with higher SNR comparatively) to evaluate the response of the proposed model to the different SNRs and applied all baseline models to them. The evaluation results of the proposed model for these synthetic datasets are summarized in Tables IV and VI. The best results are marked in bold text and underlined values represent the second-best score. Some baseline models generated empty graphs, so we placed a "NE" for those methods in the comparison table instead of mentioning the zero (0) F1 score and FDR as 1. By observing evaluation results, we can see that the performance of the proposed model increases with the increase of SNRs. From Table IV we can see that the proposed method achieved the best SHD for two datasets. Except for the zero (0.00) FDR of TCDF method for second dataset and DAG-GNN method for third dataset, the proposed method yielded the best F1 score with the lowest FDR, which means the predicted edges are mostly true. From the evaluation results of the predicted full temporal causal graphs shown in Table VI, we can summarize that the proposed method generated comparative results for synthetic datasets with different SNR.

The non-stationarity property analysis of the TKE and Arctic Sea Ice data revealed that all variables of the TKE data are non-stationary and there is a good mixture of non-stationary and stationary variables in the Arctic Sea Ice data. Our proposed model demonstrated good performance for both of these real world non-stationarity datasets (Section V-D). The simulated datasets qualify the impact of various noise distributions on the proposed model's performance. For synthetic datasets 1 and 2, respectively, we employed Gaussian and Poisson noise distributions. The evaluation findings confirm that our model can function with both types of noise (Sections V-C and V-B) and is not restricted to only these two types of noise. The comprehensive experiments with different cases and the evaluation results confirm the robustness of the proposed model for causal discovery for time series data with diverse properties.

VI. RELATED WORKS

Causal Convolutional Neural Network (CausalCNN). Recently, some studies have combined the similarities in hierarchical structures within convolutional neural networks (CNNs) and causal graphs to develop causal networks using CNNs to serve a variety of purposes. [44] proposed CexCNN - an explanation technique for CNNs using Pearl's theory of counterfactual reasoning to improve predictability. [45] proposed a Knowledge-oriented Convolutional Neural Network (K-CNN) for causal relation extraction in texts for natural language processing tasks. [46] use attribute pairs to pre-train CNN models for determining pair-wise causal direction in data. [47] proposed a novel Recurrent Convolutional Network (RCN) where they employ recurrence to 3D CNN for causal processing of videos. Earlier, [48] proposed an attention-based CNN approach for time series causal discovery, however, their approach had several limitations, including the stationarity assumption in data. It is important to note that our proposed

TABLE VI
COMPARISON OF THE PREDICTED FULL CAUSAL GRAPH BY DIFFERENT METHODS FOR SYNTHETIC DATASETS WITH DIFFERENT SNR.

METHOD	SYNTHTIC-1.1 SNR - 0.26			SYNTHTIC-1.2 SNR - 0.50			SYNTHTIC-1.3 SNR - 0.62		
	SHD	F1	FDR	SHD	F1	FDR	SHD	F1	FDR
PCMCI [13]	50	0.19	0.89	54	0.20	0.88	61	0.18	0.89
PCMCI+ [14]	46	0.20	0.88	56	0.17	0.90	39	0.23	0.86
NOTEARS-MLP [11]	26	0.31	0.80	23	0.37	0.76	27	0.30	0.81
NTS-NOTEARS [15]	NE	NE	NE	NE	NE	NE	**7**	0.25	0.50
DAG-GNN [12]	NE	NE	NE	NE	NE	NE	10	0.00	1.00
DYNOTEARS [8]	24	0.20	0.86	23	0.20	0.86	23	0.34	0.78
TCDF [48]	**10**	NE	1.00	**10**	0.16	**0.66**	10	0.28	0.60
PROPOSED	12	**0.33**	**0.72**	11	**0.42**	**0.66**	8	**0.60**	**0.45**

approach is distinguishable from these methods in both functionality and usage.

Causal Discovery on Nonlinear and Non-Stationary Data. Statistical causal discovery methods, like Granger Causality (GC), cannot handle non-linearity or non-stationarity in data. While some methods extend the traditional causal discovery method to handle non-linearity [33], [49], some approaches utilize neural network-for these extensions [6], [9]. Further, some recent research has made inroads to propose techniques applicable to non-stationary time-series. [19] proposed Constraint-based causal Discovery from NOnstationary Data (CD-NOD). Introducing a non-parametric principled framework, they demonstrated the efficiency of their approach in forecasting non-stationary data. [31] proposed a probabilistic deep learning approach, State-Dependent Causal Inference (SDCI), to perform causal discovery in conditionally stationary time-series data. [29] introduced a functional causal model (FCM) based approach for causal learning in non-stationary data with general nonlinear relationships. [32] proposed a constraint-based causal discovery approach for autocorrelated and non-stationary time series data (CDANs) to identify lagged and instantaneous causal relationships in autocorrelated and non-stationary time series data. Most recently, [30] combined Linear Non-Gaussian Acyclic Model (LiNGAM) and the Just-In-Time (JIT) framework to propose a new causal discovery method JIT-LiNGAM to identify causal relations in non-linear and non-stationary data. Recent work also includes deep learning-based methods which infer the causal graph directly from observational data using transformers, generative modeling, and adversarial learning techniques [7], [12], [48], overcoming the non-linearity constraint whereas the challenge of handling non-stationarity still prevails in them.

Limitations of Related Works. While these existing methodologies have significantly contributed to the field of time series causal discovery, several challenges persist. The inability to handle high-dimensional non-stationary datasets and the difficulty in distinguishing causation from correlation remain active research areas. Additionally, many methods struggle with interpreting and providing actionable insights, especially in complex real world applications such as climate science. By addressing the challenges posed by traditional and machine learning based approaches, our method aims to provide a comprehensive solution for effective time series causal discovery.

VII. CONCLUSION

We propose TS-CausalNN, a score-based causal structure learning method for non-linear and non-stationary time series data using a custom 2D causal convolution layer. The proposed method utilizes the power of the 2D neural networks to learn the contemporaneous and time-lagged causal relationships of all temporal variables simultaneously. The theory and computational procedure of the method are established from the continuous formulation of the acyclicity constraint of the causal DAG adapted from Zheng et al. [10]. The parallel causal convolution layer block of the proposed model keeps the causal contributor of each target/effect variable segregated from other target variables, which gives better causal stability to the generated DAG structure. The proposed model is very simple and user-friendly, as any prior knowledge of variable independence or the data generation model is not required and also, there are no assumptions regarding error distribution. We conducted extensive experiments on synthetic and real world complex time series datasets to demonstrate the performance of the proposed causal discovery model. Based on the empirical evaluation results, the proposed model demonstrates superior causal graph learning capability compared to the baseline methods.

The code is available at http://github.com/TS-CausalNN.

ACKNOWLEDGMENT

This work is supported by NSF grants: CAREER: Big Data Climate Causality (OAC-1942714) and HDR Institute: HARP - Harnessing Data and Model Revolution in the Polar Regions (OAC-2118285).

REFERENCES

[1] J. C. Rajapakse and J. Zhou, *Learning effective brain connectivity with dynamic Bayesian networks*, Neuroimage 37, 3 (2007).
[2] J. Pearl, *Probabilistic reasoning in intelligent systems - networks of plausible inference*, In Morgan Kaufmann series in representation and reasoning (1991).
[3] J. Pearl, *Models, reasoning and inference*, Cambridge, UK: Cambridge University Press 19 (2000).
[4] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, and J. Muñoz-Marí, et al., *Inferring causation from time series in Earth system sciences*, Nature communications 10 (2019).

[5] P. Spirtes, C. N Glymour, and R. Scheines. *Causation, prediction, and search*, MIT Press, Cambridge, MA, USA, 2000.

[6] S. Absar, Y. Wu, and L. Zhang, *Neural Time-Invariant Causal Discovery from Time Series Data*, In International Joint Conference on Neural Networks (2023).

[7] D. Kalainathan, O. Goudet, I. Guyon, D. Lopez-Paz, and M. Sebag, *Sam: Structural agnostic model, causal discovery and penalized adversarial learning*, (2018).

[8] R. Pamfil, N. Sriwattanaworachai, S. Desai, P. Pilgerstorfer, K. Georgatzis, P. Beaumont, and B. Aragam, *Dynotears: Structure learning from time-series data*, In International Conference on Artificial Intelligence and Statistics (2020).

[9] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. B. Fox, *Neural granger causality*, IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 8 (2021).

[10] X. Zheng, B. Aragam, P. Ravikumar, and E. P. Xing, *DAGs with NO TEARS: Continuous Optimization for Structure Learning*, In Advances in Neural Information Processing Systems (2018).

[11] X. Zheng, C. Dan, B. Aragam, . Ravikumar, and E. P. Xing,*Learning sparse nonparametric DAGs*, In International Conference on Artificial Intelligence and Statistics (2020).

[12] Y. Yu, J. Chen, T. Gao, and M. Yu, *DAG-GNN: DAG structure learning with graph neural networks*, In International Conference on Machine Learning (2019).

[13] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, *Detecting and quantifying causal associations in large non-linear time series datasets*, Science Advances 5, 11 (2019).

[14] J. Runge., *Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets*, In the Uncertainty in Artificial Intelligence (UAI) (2020).

[15] X. Sun, O. Schulte, G. Liu, and P. Poupart,*NTS- NOTEARS: Learning Nonparametric DBNs With Prior Knowledge*, In International Conference on Artificial Intelligence and Statistics (2023).

[16] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*, The MIT Press, Cambridge, MA, USA (2017).

[17] R. D. Shah and J. Peters, *The hardness of conditional independence testing and the generalised covariance measure*, The Annals of Statistics 48, 3 (2020).

[18] D. Entner and P. O. Hoyer, *On causal discovery from time series data using FCI*, Probabilistic graphical models, 2010.

[19] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. *Causal discovery from heterogeneous nonstationary data*, The Journal of Machine Learning Research 21, 1 (2020).

[20] A. Gerhardus and J. Runge, *High-recall causal discovery for autocorrelated time series with latent confounders*, Advances in Neural Information Processing Systems 33 (2020).

[21] J. O. Hinze, *Turbulence*, McGraw-Hill (1975).

[22] W. C. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, Z. Liu, J. Berner, W. Wang, J. G. Powers, M. G. Duda, D. M. Barker, et al., *A description of the advanced research WRF version 4*, NCAR tech. note ncar/tn-556+ str 145 (2019).

[23] W. I. Gustafson, A. M. Vogelmann, Z. Li, X. Cheng, K. K. Dumas, S. Endo, K. L. Johnson, B. Krishna, T. Fairless, and H. Xiao, *The large-eddy simulation (LES) atmospheric radiation measurement (ARM) symbiotic simulation and observation (LASSO) activity for continental shallow convection*, Bulletin of the American Meteorological Society 101, 4 (2020).

[24] S. Endo, A. M. Fridlind, W. Lin, A. M. Vogelmann, T. Toto, A. S. Ackerman, G. M. McFarquhar, R. C. Jackson, H. H. Jonsson, and Y. Liu, *RACORO continental boundary layer cloud investigations: 2. Large-eddy simulations of cumulus clouds and evaluation with in situ and ground-based observations*, Journal of Geophysical Research: Atmospheres 120, 12 (2015).

[25] A. D. Sanford and I. A. Moosa, *A Bayesian network structure for operational risk modelling in structured finance operations*, Journal of the Operational Research Society 63 (2012).

[26] M. O. Appiah, *Investigating the multivariate Granger causality between energy consumption, economic growth and CO2 emissions in Ghana*, Energy Policy 112 (2018).

[27] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*, MIT press (2009).

[28] D. E. Heckerman, E. J. Horvitz, and B. N. Nathwani, *Toward normative expert systems: Part i the pathfinder project*, Methods of information in medicine 31, 02 (1992).

[29] T. Wu, X. Wu, X. Wang, S. Liu, and H. Chen, *Nonlinear Causal Discovery in Time Series*, In the 31st ACM International Conference on Information and Knowledge Management (2022).

[30] D. Fujiwara, K. Koyama, K. Kiritoshi, T. Okawachi, T. Izumitani, and S. Shimizu, *Causal Discovery for Non-stationary Non-linear Time Series Data Using Just-In-Time Modeling*, In the Second Conference on Causal Learning and Reasoning (2023).

[31] C. B. Rodas, R. Tu, and H. Kjellström, *Causal discovery from conditionally stationary time-series*, arXiv preprint arXiv:2110.06257 (2021).

[32] M. H. Ferdous, U. Hasan, and M. O. Gani, *CDANs: Temporal Causal Discovery from Autocorrelated and Non-Stationary Time Series Data*, In Machine Learning for Healthcare Conference (2023).

[33] M. T. Bahadori and Y. Liu, *On causality inference in time series*, In AAAI Fall Symposium Series (2012).

[34] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al., *Recent advances in convolutional neural networks*, Pattern recognition 77 (2018).

[35] 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 4 (2021).

[36] M. C. Serreze and J. Stroeve, *Arctic sea ice trends, variability and implications for seasonal ice forecasting*, In Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (2015).

[37] I. Simmonds, *Comparing and contrasting the behaviour of Arctic and Antarctic sea ice over the 35 year period 1979-2013*, Annals of Glaciology (2015).

[38] J. Cohen, J. A. Screen, J. C. Furtado, M. Barlow, D. Whittleston, D. Coumou, J. Francis, K. Dethloff, D. Entekhabi, J. Overland, et al., *Recent Arctic amplification and extreme mid-latitude weather*, Nature geoscience 7, 9 (2014).

[39] I. Simmonds and P. D. Govekar, *What are the physical links between Arctic sea ice loss and Eurasian winter climate?*, Environmental Research Letters 9, 10 (2014).

[40] Y. Yao, D. Luo, A. Dai, and I. Simmonds, *Increased quasi stationarity and persistence of winter Ural blocking and Eurasian extreme cold events in response to Arctic warming. Part I: Insights from observational analyses*, Journal of Climate 30, 10 (2017).

[41] D. Luo, X. Chen, A. Dai, and I. Simmonds,*Changes in atmospheric blocking circulations linked with winter Arctic warming: A new perspective*, Journal of Climate 31, 18 (2018).

[42] D. Luo, X. Chen, J. Overland, I. Simmonds, Y. Wu, and P. Zhang, *Weakened potential vorticity barrier linked to recent 1127 winter Arctic sea ice loss and midlatitude cold extremes*, Journal of Climate 32, 1128 14 (2019).

[43] U. Hasan, E. Hossain, and M. O. Gani, *A survey on causal discovery methods for iid and time series data*, Transactions on Machine Learning Research (2023).

[44] H. Debbi, *Causal explanation of convolutional neural networks*, Machine Learning and Knowledge Discovery in Databases, ECML PKDD (2021).

[45] P. Li, and K. Mao,*Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts*, Expert Systems with Applications (2019).

[46] K. Singh, G. Gupta, L. Vig, G. Shroff, and P. Agarwal, *Deep convolutional neural networks for pairwise causality*, arXiv preprint arXiv:1701.00597 (2017).

[47] G. Singh, and F. Cuzzolin, *Recurrent convolutions for causal 3D CNNs*, In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019).

[48] M. Nauta, D. Bucur, and C. Seifert, *Causal discovery with attention-based convolutional neural networks*, Machine Learning and Knowledge Extraction 1, no. 1 (2019).

[49] A. Gerhardus, and J. Runge, *High-recall causal discovery for autocorrelated time series with latent confounders*, Advances in neural information processing systems 33 (2020).