MXMap: A Multivariate Cross Mapping Framework for Causal Discovery in Dynamical Systems

Elise Zhang   ELISE.ZHANG@MAIL.MCGILL.CA
McGill University, Montréal, QC Canada

François Mirallès   MIRALLES.FRANCOIS@HYDROQUEBEC.COM
Hydro-Québec Research Institute (IREQ), Varennes, QC Canada

Raphaël Rousseau-Rizzi   ROUSSEAU-RIZZI.RAPHAEL@HYDROQUEBEC.COM
Hydro-Québec Research Institute (IREQ), Varennes, QC Canada

Di Wu   DI.WU5@MCGILL.CA
McGill University, Montréal, QC Canada

Arnaud Zinflou   ZINFLOU.ARNAUD@HYDROQUEBEC.COM
Hydro-Québec Research Institute (IREQ), Varennes, QC Canada

Benoit Boulet   BENOIT.BOULET@MCGILL.CA
McGill University, Montréal, QC Canada

Editors: Biwei Huang and Mathias Drton

Abstract

Convergent Cross Mapping (CCM) is a powerful method for detecting causality in coupled non-linear dynamical systems, providing a model-free approach to capture dynamic causal interactions. Partial Cross Mapping (PCM) was introduced as an extension of CCM to address indirect causality in three-variable systems by comparing cross-mapping quality between direct cause-effect mapping and indirect mapping through an intermediate conditioning variable. However, PCM remains limited to univariate delay embeddings in its cross-mapping processes. In this work, we extend PCM to the multivariate setting, introducing multiPCM, which leverages multivariate embeddings to more effectively distinguish indirect causal relationships. We further propose a multivariate cross-mapping framework (MXMap) for causal discovery in dynamical systems. This two-phase framework combines (1) pairwise CCM tests to establish an initial causal graph and (2) multiPCM to refine the graph by pruning indirect causal connections. Through experiments [1] on simulated data and the *ERA5 Reanalysis* weather dataset, we demonstrate the effectiveness of MXMap. Additionally, MXMap is compared against several baseline methods, showing advantages in accuracy and causal graph refinement.

Keywords: Causal inference, state-space reconstruction, convergent cross-mapping, partial cross mapping, nonlinear dynamical system

---

1. Implementation at https://github.com/elisejiuqizhang/multiPCM.

1. Introduction

Nonlinear dynamical systems are omnipresent across scientific disciplines, and understanding causal relationships in these systems is crucial for unveiling the underlying mechanisms that drive system behaviors. Classic causal inference methods, such as Granger Causality (GC) (Granger, 1969) and other functional causal models (FCMs), including the Additive Noise Model (ANM) (Hoyer et al., 2008; Liu et al., 2024) and the Post Nonlinear Model (PNL) (Zhang et al., 2015; Keropyan et al., 2023), struggle with these systems due to their assumption of a predictive relationship from cause to effect, which does not hold in the presence of complex dynamics like coupling and chaos.

Convergent Cross Mapping (CCM) (Sugihara et al., 2012; Barraquand et al., 2021) was proposed as a model-free approach for bivariate causal inference in coupled dynamical systems. CCM addresses these limitations by leveraging state-space manifold reconstructions and cross mapping between reconstructed embeddings. Since its introduction, CCM has inspired further developments, including Partial Cross Mapping (PCM) (Leng et al., 2020), which aims to distinguish indirect from direct causalities in three-variable systems. However, PCM is limited to mapping operations between univariate delay embeddings, which can be less effective or even fail when dealing with complex systems with multiple interconnected variables (Chen et al., 2022).

To overcome this limitation, we propose **multiPCM**, an extension of PCM to the multivariate setting that allows for more effective causal inference by utilizing cross mapping via multivariate embeddings. We further integrate multiPCM with bivariate CCM into a two-phase framework named **MXMap** (Multivariate Cross Mapping for Causal Discovery). The proposed framework is designed for multivariate causal discovery, and is not only confined to assumptions of directed acyclic graphs (DAGs) but can also handle cycles. In the first phase, bivariate CCM generates an initial, potentially dense causal graph; In the second phase, multiPCM prunes indirect connections, refining the graph to isolate direct causal relationships. We systematically evaluate multiPCM and MXMap on benchmark datasets, including both simulated ecosystems and real-world meteorological data.

The contributions of this work are summarized as follows:

- **Extension of PCM to multivariate settings**: We introduce multiPCM, which extends PCM to utilize multivariate delay embeddings for more robust causal inference in high-dimensional dynamical systems.

- **Two-phase causal discovery framework**: We propose MXMap, combining bivariate CCM with multiPCM to generate and refine causal graphs in nonlinear dynamical systems, which can also detect cycles.

- **Comprehensive evaluation on nonlinear dynamical systems**: We validate multiPCM and MXMap on simulated and real-world datasets. MXMap is compared against multiple baseline methods — including tsFCI, VAR-LiNGAM, PCMCI, Granger Causality, DYNOTEARS, SLARAC — demonstrating advantages in accuracy and refinement capabilities.

2. Preliminaries

2.1. State-Space Reconstruction (SSR)

In physical continuous-time dynamical systems, the interplay between driving forces and dissipation leads systems to settle into characteristic behaviors, represented by attractor manifolds in state

space (Milnor, 1985). Understanding these attractors is crucial for interpreting the system's dynamics and predicting future behavior. However, real-world measurements are often limited, making it infeasible to observe all variables required to fully characterize the state-space attractor.

State-Space Reconstruction (SSR) (Vlachos and Kugiumtzis, 2008) addresses this challenge: *Given an n-dimensional observed time series from an N-dimensional dynamical system ($n < N$), can we recover the attractor manifold, and thereby the higher-dimensional dynamics, from the lower-dimensional observations?* A common approach is to use sequences of lagged observations to reconstruct a delay embedding (DE) that approximates the system's attractor. Whitney's Embedding Theorem (Whitney, 1936) and Takens' Embedding Theorem (Takens, 2006) establish that this reconstruction is diffeomorphic (i.e., a continuously differentiable and invertible mapping) to the true attractor under certain conditions (Sauer et al., 1991). When these conditions are satisfied, such delay embeddings are termed "shadow manifolds" and serve as low-dimensional approximations of the system (Sugihara et al., 2012).

Following (Vlachos and Kugiumtzis, 2010; Butler et al., 2023), we introduce the univariate delay-coordinate embedding used in Takens' Theorem. Suppose an attractor $\mathcal{A}$ exists for the dynamical system, and a time series $\{x_t\}_{t=0}^T$ is observed from one state variable **x** sampled at a constant rate. Given delay τ and embedding dimension E, where τ and E are positive integers, the vector signal $\vec{\mathbf{m}}_x(t)$ of lagged values is defined as:

$$\vec{\mathbf{m}}_x(t) := \left[x_t, x_{t-\tau}, x_{t-2\tau}, x_{t-3\tau}, \ldots, x_{t-(E-2)\tau}, x_{t-(E-1)\tau} \right] \tag{1}$$

As time progresses, these vectors form an E-dimensional delay embedding $\mathbf{M}_x$. The lag τ determines the observation time scale for reconstruction, while the embedding dimension E defines the complexity of the embedding. Takens' theorem suggests that E should be greater than twice the fractal dimension of the attractor $\mathcal{A}$, i.e., $E > 2 \cdot \dim(\mathcal{A})$ (Sauer et al., 1991; Kugiumtzis, 1996). In practice, τ and E are determined empirically. Time-delayed autocorrelation (Kugiumtzis, 1996) and delay mutual information (Fraser and Swinney, 1986; Kliková and Raidl, 2011) are commonly used to select an optimal τ, while the *false nearest neighbors* (FNN) method (Kennel et al., 1992) is typically used to determine E, by tracking changes in nearest neighbors as embedding dimensions increase.

2.2. Convergent Cross Mapping (CCM)

Causality in a discrete-time dynamical system (Butler et al., 2023; Cummins et al., 2015) can be defined as follows: given two state variables **x** and **y**, if the future evolution of **y** depends on **x**, then **x** is said to cause **y**, denoted as $\mathbf{x} \Rightarrow \mathbf{y}$. This causal influence can be represented in a state-space equation, as shown in Eq. 2:

$$\mathbf{y}_{t+1} = \mathcal{F}_y \left(\mathbf{y}_t, \mathbf{x}_t \right) \tag{2}$$

The relationship between **x** and **y** can be unidirectional ($\mathbf{x} \Rightarrow \mathbf{y}$ or $\mathbf{y} \Rightarrow \mathbf{x}$), bidirectional ($\mathbf{x} \Leftrightarrow \mathbf{y}$), or there may be no causal link at all.

CCM (Sugihara et al., 2012) leverages the diffeomorphism between reconstructed shadow manifolds, as stated in Takens' Theorem. Cross mapping measures how well local neighborhoods in one reconstructed manifold map to the corresponding neighborhoods in another. For delay embeddings $\mathbf{M}_x$ and $\mathbf{M}_y$ reconstructed from **x** and **y**, if $\mathbf{M}_x$ and $\mathbf{M}_y$ are both valid shadow manifolds of the attractor $\mathcal{A}$, they are diffeomorphic to each other via their relationship to $\mathcal{A}$.

If $\mathbf{x}$ causes $\mathbf{y}$ ($\mathbf{x} \Rightarrow \mathbf{y}$), observations of $\mathbf{y}$ should contain information about $\mathbf{x}$. This allows the reconstruction of the dynamics of $\mathbf{x}$ from $\mathbf{y}$, but not necessarily vice versa. In this case, the quality of mapping from $\mathbf{M}_y$ to $\mathbf{M}_x$ should be better compared to the reverse direction, indicating a causal link from $\mathbf{x}$ to $\mathbf{y}$.

CCM uses a k-nearest neighbor (kNN) regression approach (also known as *simplex projection*) to evaluate the quality of cross mapping. Given time series $\{x_t\}_{t=0}^T$ and $\{y_t\}_{t=0}^T$, to verify whether $\mathbf{x} \Rightarrow \mathbf{y}$, the procedure is as follows:

[1] Construct delay embeddings $\mathbf{M}_x, \mathbf{M}_y$ with appropriate delay τ and embedding dimension E.

[2] For each point in $\mathbf{M}_y$, identify the k-nearest neighbors $\mathcal{N}_y$.

[3] Use the timestamps of $\mathcal{N}_y$ to find corresponding points $\hat{\mathcal{N}}_x$ on $\mathbf{M}_x$ and compute a weighted average to form a projected reconstruction $\hat{\mathbf{M}}_x$, hence the reconstructed $\hat{\mathbf{x}}$.

[4] Calculate the correlation score $\rho_{x \Rightarrow y}$ between the true $\mathbf{x}$ and the reconstructed $\hat{\mathbf{x}}$.

[5] Repeat these steps with increasing sequence length; if $\mathbf{x} \Rightarrow \mathbf{y}$, the correlation score should converge, indicating a valid cross map.

The same procedure is repeated for the reverse causal assumption to yield another correlation score $\rho_{y \Rightarrow x}$. The correlation scores $\rho_{x \Rightarrow y}$ and $\rho_{y \Rightarrow x}$ quantify the cross mapping quality, where a higher score in one direction suggests a stronger causal link. In practice, if the input length L is large enough, we consider that the yielded correlation scores are already in the convergence zone, and can be used as final correlation estimates.

2.3. Partial Cross Mapping (PCM)

The original CCM does not distinguish between direct and indirect causality. For three variables $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ in a causal chain ($\mathbf{x} \Rightarrow \mathbf{y} \Rightarrow \mathbf{z}$), CCM may incorrectly identify a direct causal link between $\mathbf{x}$ and $\mathbf{z}$ due to transitivity through $\mathbf{y}$. Partial Cross Mapping (PCM) (Leng et al., 2020; Jiang et al., 2023) was proposed to distinguish between direct and indirect causal relationships. In a causal chain like $\mathbf{x} \Rightarrow \mathbf{y} \Rightarrow \mathbf{z}$, PCM aims to determine whether the causal link between $\mathbf{x}$ and $\mathbf{z}$ is direct or mediated by $\mathbf{y}$.

A PCM test considers three variables: the *potential cause* $\mathbf{x}$, the *condition* $\mathbf{y}$, and the *potential effect* $\mathbf{z}$. The goal is to assess whether there is a direct link between $\mathbf{x}$ and $\mathbf{z}$. This is done by performing cross mapping between the shadow manifolds of each variable to obtain a reconstruction of $\mathbf{x}$, denoted by $\hat{\mathbf{x}}^{\mathbf{z}}$ (from $\mathbf{z}$ to $\mathbf{x}$), and another reconstruction of $\mathbf{x}$ via $\mathbf{y}$, denoted by $\hat{\mathbf{x}}^{\hat{\mathbf{y}}^{\mathbf{z}}}$ (first from $\mathbf{z}$ to $\mathbf{y}$, then from $\mathbf{y}$ to $\mathbf{x}$). The correlation scores are defined as follows:

$$\rho_{All} = |\text{Corr}(\mathbf{x}, \hat{\mathbf{x}}^{\mathbf{z}})| \qquad\qquad \rho_{Direct} = \left| \text{ParCorr}(\mathbf{x}, \hat{\mathbf{x}}^{\mathbf{z}} | \hat{\mathbf{x}}^{\hat{\mathbf{y}}^{\mathbf{z}}}) \right| \qquad (3)$$

ρ_{All} represents the correlation between the original $\mathbf{x}$ and the reconstruction $\hat{\mathbf{x}}^{\mathbf{z}}$, capturing apparent information transfer through all paths. ρ_{Direct} represents the partial correlation, conditioning on the intermediate variable $\mathbf{y}$ to assess direct information transfer between $\mathbf{x}$ and $\mathbf{z}$. If no direct causal link exists, the direct information transfer should be significantly reduced after conditioning on $\mathbf{y}$.

PCM uses an empirical threshold $H \in [0, 1)$ to determine causality:

- If $\rho_{All} \geq \rho_{Direct} \geq H$, a direct causal link from **x** to **z** is inferred.

- If $\rho_{All} \geq H \gg \rho_{Direct}$, only indirect causality is suggested.

- If $H > \rho_{All} \geq \rho_{Direct}$, no causal relationship is inferred.

To distinguish direct from indirect links, we propose an adaptation in our work, to use the correlation ratio γ as an alternative:

$$\gamma = \frac{\rho_{Direct}}{\rho_{All}} \tag{4}$$

A smaller ratio γ implies negligible direct information transfer after conditioning, suggesting indirect causality via **y**. Conversely, a larger γ indicates strong direct information transfer, suggesting a direct causal link. An empirical ratio threshold $\gamma^* \in (0, 1)$ is used to decide whether to retain or eliminate the direct link based on how important such causal influence is.

3. Methodology

We present key components of our approach: multivariate state space reconstruction for capturing coupled system dynamics; multivariate partial cross mapping to distinguishes between direct and indirect causality in multivariate settings; MXMap, the proposed framework leveraging the two for multivariate causal discovery.

3.1. Multivariate State Space Reconstruction (multiSSR)

Consider a set of K observed time series denoted as $\left\{ x_t^{[k]} \right\}_{t=0}^{T}, k = 1, 2, 3, ..., K$ from a dynamical system. Building upon the univariate delay embedding introduced in Section 2.1, each time series generates a univariate embedding $\mathcal{M}_X^{[k]}$. The vector $\vec{\mathbf{m}}_x^{[k]}(t)$ on $\mathcal{M}_x^{[k]}$ is given by:

$$\vec{\mathbf{m}}_x^{[k]}(t) = \left[x_t^{[k]}, x_{t-\tau}^{[k]}, x_{t-\tau \cdot 2}^{[k]}, x_{t-\tau \cdot 3}^{[k]}, ..., x_{t-\tau \cdot (E-2)}^{[k]}, x_{t-\tau \cdot (E-1)}^{[k]} \right] \tag{5}$$

Multivariate State-Space Reconstruction (multiSSR) constructs a multivariate embedding by stacking all K univariate delay embeddings. A vector $\mathbf{m}_x(t)$ on $\mathcal{M}_x(t)$ is represented as:

$$\vec{\mathbf{m}}_x(t) = \left[\vec{\mathbf{m}}_x^{[1]}(t) \quad \vec{\mathbf{m}}_x^{[2]}(t) \quad \vec{\mathbf{m}}_x^{[3]}(t) \quad ... \quad \vec{\mathbf{m}}_x^{[K]}(t) \right] \tag{6}$$

For simplicity and consistency, we adopt a uniform multiSSR scheme, where the delay values τ and embedding dimensions E are the same for all K time series, as suggested by (Vlachos and Kugiumtzis, 2010).

3.2. Multivariate Partial Cross Mapping (multiPCM)

The original PCM method considers three univariate inputs: the potential cause, effect, and condition variables. Multivariate Partial Cross Mapping (multiPCM) extends this by allowing the condition set to be multivariate, making it better suited for high-dimensional systems.

Consider a set of variables $\mathcal{G}$, with a quasi-chain structure $X_1 \Rightarrow \mathcal{G} \setminus \{X_1, X_2\} \Rightarrow X_2$. Here, X_1 is the alleged cause, X_2 is the alleged effect, and the rest of the variables, denoted as *Conds* := $\mathcal{G} \setminus \{X_1, X_2\}$, serve as the condition set.

The two univariate inputs, X_1 and X_2, form univariate delay embeddings, while the multivariate condition set forms a multivariate embedding as described in Section 3.1.

Similar to the original PCM as described in Section 2.3, we conduct the apparent (univariate) and conditioned (multivariate) cross mappings to obtain the two correlation scores.

- **Apparent cross mapping:** Perform univariate cross mapping to reconstruct X_1 from X_2, denoted as $\hat{X}_1^{X_2}$.

- **Conditioned cross mapping:** Reconstruct the condition set from X_2, denoted by $\widehat{Conds}^{X_2}$, and then reconstruct X_1 from $\widehat{Conds}^{X_2}$, resulting in $\hat{X}_1^{\widehat{Conds}^{X_2}}$.

The two correlation scores are then computed as follows:

$$\rho_{All} = \left| Corr(X_1, \hat{X}_1^{X_2}) \right| \qquad \rho_{Direct} = \left| ParCorr(X_1, \hat{X}_1^{X_2} | \hat{X}_1^{\widehat{Conds}^{X_2}}) \right| \qquad (7)$$

Following a similar reasoning as in Section 2.3, we determine whether a direct causal link exists between X_1 and X_2 by calculating the correlation ratio $\gamma = \frac{\rho_{Direct}}{\rho_{All}}$ and comparing it against an empirical threshold.

3.3. Multivariate Cross Mapping (MXMap) for Multivariate Causal Discovery in Dynamical Systems

We propose MXMap as a two-phase framework for multivariate causal discovery in dynamical systems. Inspired by the structure of RESIT (Regression with Subsequent Independence Test) (Peters et al., 2014) algorithm, MXMap first establishes an initial causal graph and then prunes indirect connections (See Algorithm 1 in Appendix A).

The details of each phase are as follows:

- **Phase 1 (Establish Initial Causal Graph):** In the first phase, MXMap applies exhaustive bivariate CCM tests between all pairs of variables to establish an initial causal graph. For each pair of variables, we embed their time series using delay embedding parameters (τ, E) and compute delay embeddings $\mathcal{M}_{X_i}$ and $\mathcal{M}_{X_j}$. Cross mapping is then performed between these embeddings to reconstruct each variable, yielding reconstructed series $\hat{X}_i$ and $\hat{X}_j$. Correlation scores are calculated for both directions—$\beta_{X_i \Rightarrow X_j}$ and $\beta_{X_j \Rightarrow X_i}$. The direction with the higher score is chosen as the causal direction, and a link is added to the initial graph accordingly. For this CCM phase, if both scores are below a correlation threshold (0.5), no link is established.

- **Phase 2 (Prune Indirect Causal Connections):** The initial causal graph generated in Phase 1 may contain indirect connections, which can result in an overly dense graph. In the second phase, MXMap uses multiPCM to refine the graph by pruning indirect links. For each parent-child pair in the initial graph, we identify all intermediate variables forming paths between them. We then apply multiPCM to assess whether the causal link between the parent and child is direct or goes through intermediate variables. Specifically, we embed the parent, child, and intermediate variables, and compute two correlation scores: ρ_{All}, representing overall correlation, and ρ_{Direct}, representing partial correlation conditioned on the intermediate

6

variables and assessing the direct information transfer between parent and child variables. We calculate the correlation ratio $\gamma = \frac{\rho_{Direct}}{\rho_{All}}$ and compare it to a predefined ratio threshold $\gamma^\star$. If ratio γ is below $\gamma^\star$, the direct information transfer isn't considered important enough, and the parent-child link is considered indirect and is removed from the graph.

4. Experiments

We evaluate the proposed multiPCM and MXMap on simulated and real-world dynamical systems as described in the Section 4.1 below. MXMap performance is then compared with several established multivariate causal inference methods), including RESIT (Peters et al., 2014), tsFCI (Entner and Hoyer, 2010), VAR-LiNGAM (Hyvärinen et al., 2010), PCMCI (Runge et al., 2019), Granger Causality (Granger, 1969), DYNOTEARS (Pamfil et al., 2020), and SLARAC (Weichwald et al., 2020). Overview of these methods and how the outputs of each model are interpreted can be found in Appendix F.

For RESIT and VAR-LiNGAM, we adopt python implementations in the `LiNGAM` library (Shimizu, 2014); tsFCI is adapted based on the implementation of FCI from the `causal-learn` library (Zheng et al., 2024); PCMCI is from the `tigramite` library (Runge et al., 2023); Granger Causality is also from `causal-learn`; DYNOTEARS is implemented using the `Causalnex` (Beaumont et al., 2021) library; SLARAC is from the `tidybench` repository (Weichwald et al., 2020). Experimental setup is provided in Appendix G.

4.1. Data

4.1.1. SIMULATED DATA: SPECIES INTERACTION SYSTEMS

Following similar data generation scheme as in Leng et al. (2020), we generate dynamical systems of varying complexity, from 3-variable to 7-variable. These systems are derived and adapted from the Lotka-Volterra competition models (Volterra, 1931; Lotka, 1925; Roques and Chekroun, 2011) that characterize species interactions and exhibit chaotic behaviors. Example of a 3-species system is shown in Eq. 8.

$$
\begin{aligned}
x_t &= x_{t-1} \left(\alpha_x - \alpha_x x_{t-1} - \beta_{yx} y_{t-1} - \beta_{zx} z_{t-1} \right) \cdot \eta_x + \epsilon_x \\
y_t &= y_{t-1} \left(\alpha_y - \alpha_y y_{t-1} - \beta_{xy} x_{t-1} - \beta_{zy} z_{t-1} \right) \cdot \eta_y + \epsilon_y \\
z_t &= z_{t-1} \left(\alpha_z - \alpha_z z_{t-1} - \beta_{xz} x_{t-1} - \beta_{yz} y_{t-1} \right) \cdot \eta_z + \epsilon_z
\end{aligned}
\tag{8}
$$

For the 3-species systems (Eq. 8), the coefficients α for autonomous dynamics are set respectively as $\alpha_x = 3.70, \alpha_y = 3.78, \alpha_z = 3.72$. The coupling coefficient β_{ij} is either 0 or 0.35, depending on whether the causal interaction is present or absent. For higher dimensional systems, we follow a similar logic, where the autonomous dynamics coefficients α are sampled in range $[3.70, 3.80)$, and the coupling coefficients β are 0.35 when there is causal interaction between a variable pair (0 if no causal interaction). These systems are used to evaluate the effectiveness of multiPCM and MXMap in capturing both direct and indirect causal links in controlled scenarios. The ground truth causal structures of all used systems (from 3-variable to 7-variable) are demonstrated in Appendix J.

4.1.2. ERA5 REANALYSIS METEOROLOGICAL DATA

We also evaluate MXMap on real-world meteorological data from the ERA5 Global Reanalysis dataset (Hersbach et al., 2020), provided by the Copernicus program by ECMWF. The ERA5 dataset

offers hourly climate variables over a global scale, allowing us to investigate causality in a practical environmental setting.

We extract hourly winter data (December to February) for the Montreal region from 1981 to 2023. Two setups were chosen to assess the effectiveness of causal discovery methods:

- **3V Chain**: total cloud water (tcw) $\Rightarrow$ total radiation (rad) $\Rightarrow$ near-ground temperate (T_{2m}).

- **5V System**: This system includes solar radiation (rad_{solar}), terrestrial radiation (rad_{terr}), near-ground temperature advection (T_{adv950}), total cloud water (tcw) and near-ground temperature (T_{2m}). We examine whether well-established causal relationships can be detected, namely $rad_{solar} \Rightarrow T_{2m}$, $rad_{terr} \Rightarrow T_{2m}$, $T_{adv950} \Rightarrow T_{2m}$, $tcw \Rightarrow rad_{solar}$.

Detailed explanations of these mechanisms and meteorological contexts are provided in Appendix D. By focusing on winter data, when these causal mechanisms are most pronounced, these setups serve as an ideal test case for assessing MXMap's ability to uncover causal relationships in complex, real-world environmental systems.

4.2. Validation of multiPCM

To validate the effectiveness of multiPCM in distinguishing between direct and indirect causal relationships, we perform experiments on simulated four-variable systems generated without noise. Specifically, we evaluate three scenarios: purely direct causality, purely indirect causality, and combined direct and indirect causality, as illustrated in Table 1. The causal relationships of interest are highlighted in color, while the other variables are shown in gray and grouped together to form a multivariate embedding, used as the condition set ($Conds$) for multiPCM.

We use an input length of $L = 3500$ time steps for all tests and conduct multiPCM on a range of lag and embedding dimension values ($\tau, E \in \{1, 2, 3, \ldots, 8\}$). The results of the grid search are presented in Table 1 (complete table with more cases in Appendix B), where we analyze the correlation ratio $\gamma = \rho_{Direct}/\rho_{All}$ and the predicted labels indicating whether direct causality between colored nodes is rejected and to be removed based on a PCM threshold of 0.45 (this empirical threshold selection is discussed in Appendix C). Red label indicates rejection, since only indirect causality exists; while blue label indicates the existence direct causality, hence the link between colored nodes should be kept.

The experimental results demonstrate the effectiveness of multiPCM in distinguishing direct and indirect causal relationships under different scenarios:

1. **Direct Causality:** For cases involving direct causality ($Direct$ and $Both$), we observe that both correlation scores, ρ_{All} and ρ_{Direct}, are significantly higher when the lag τ is small, and start dropping as lag increases. This trend is also consistent with observations from previous works on cross mapping in nonlinear systems.

2. **Indirect Causality:** In the case of purely indirect causality, the correlation profiles tend to be inconsistent across different lags and embedding dimensions, showing fluctuating surfaces rather than a clear decreasing trend. This behavior is characteristic of indirect interactions, as these relationships become weaker and more unstable with increasing delay.

3. **Optimal Hyperparameter Range:** The grid search results for predicted labels further illustrate the behavior of multiPCM. There exists an ideal range of lag and embedding dimension values
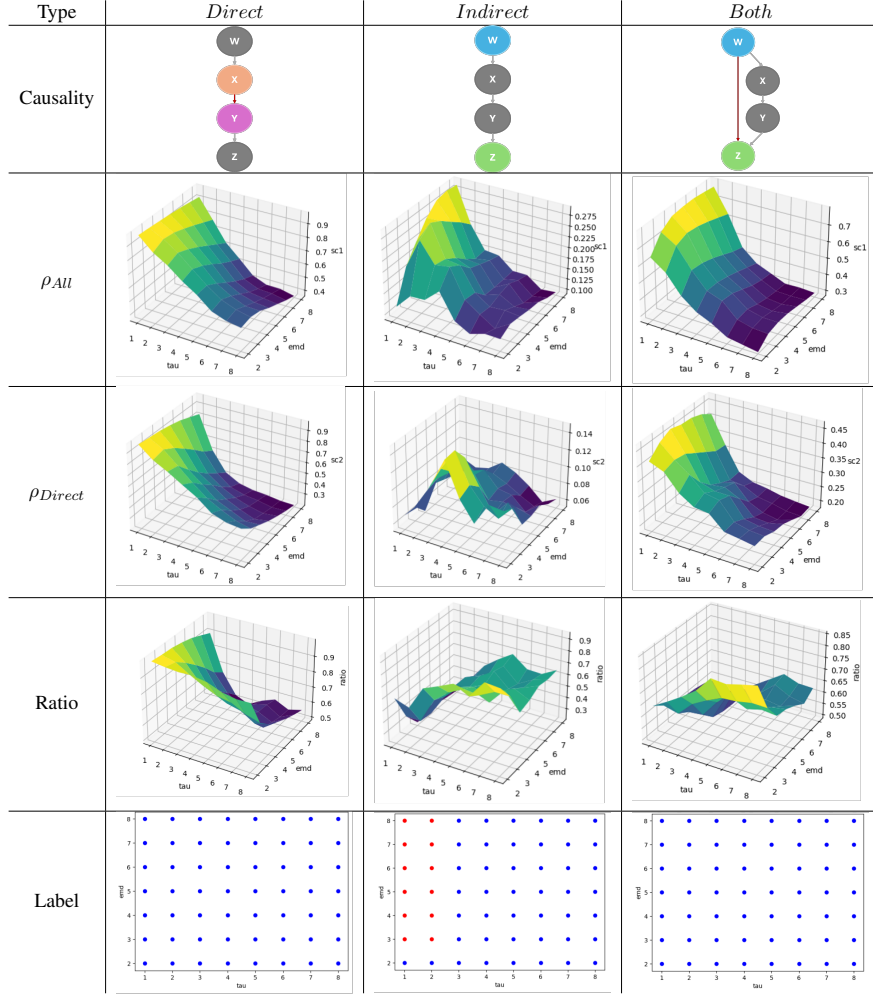
Table 1: Performance of multiPCM (Full Table in Appendix B): Profiles of correlation scores, correlation ratios, predicted label ($thres = 0.45$) under grid search. Red dot indicates there isn't direct causality between the colored nodes, while blue indicates there is direct causality between the colored nodes.

for accurately inferring causality: Here in Table 1, when the multiPCM lag (τ) is less than 2 and the embedding dimension (E) is in the interval of $[3, 8]$, multiPCM produces consistent and accurate results across these three causal structures. In practice, if the approximate system dimension and timescale of the lag are known, the appropriate selections for τ and E would likely align closely with the ground truth values: For the demonstrated 3-variable (3V) and 4-variable (4V) systems, the actual system dimension is 3 or 4, the generating dynamics use a lag of 1 (All these ground-truth τ and E values fall in the detected ranges above). Notably, the selection of E seems to be more tolerant for slight overestimation when the true state dimension is unknown. Thus, when the exact ground truth is unavailable, a slightly higher dimension for delay embedding may still yield reliable results, and can potentially help account for latent variables.

9

These observations show that multiPCM is capable of correctly distinguishing direct and indirect causalities in multivariate scenarios. By performing cross mapping with multivariate embeddings, multiPCM achieves consistent causal inference that is robust across different lag and embedding dimension configurations.

4.3. Prediction Consistency

Mirage correlations are common in nonlinear dynamical systems. Coupled nonlinear systems often exhibit transient correlations between variables, which may change or disappear entirely when different subsequences are sampled. This presents a significant challenge for causal discovery, particularly for methods relying on consistent predictive relationships, such as Granger causality. The original CCM paper Sugihara et al. (2012) illustrated this phenomenon using a bivariate system of competing species.

| (a) Subsequence 1 | (b) Subsequence 2 | (c) Subsequence 3 |
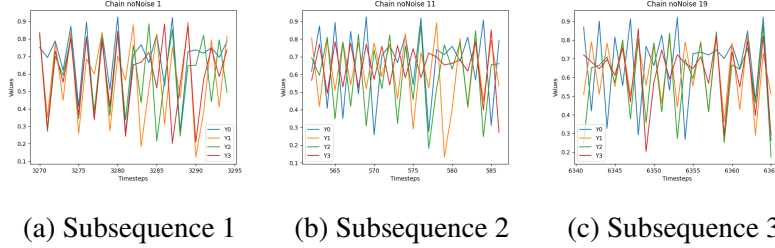
Figure 1: Visualizations of subsequences from a 4-species chain: In the same sequence, correlations between variables can be positive, negative or zero when sampling from different start points.

To evaluate the robustness of our proposed approach in such scenarios, we first illustrate the chaotic behavior and mirage correlations in the noise-free four-species chain system ($w \Rightarrow x \Rightarrow y \Rightarrow z$), as defined in Section 4.1.1. We randomly sample different starting points and visualize subsequences (of length 25) from these sampled points in Fig. 1. As shown, the correlations between variables vary widely across different subsequences, exhibiting positive, near-zero, and even negative correlations.

Model	MXMap	RESIT-MLP				
Prediction						
Precision	1.0	0.5000	0.3333	0.2500	0.5000	0.4000
Recall	1.0	1.0	0.6667	0.3333	0.6667	0.6667
F1	1.0	0.6667	0.4444	0.2857	0.5714	0.5000
SHD	0	3	5	5	3	4
Count	10	4	3	1	1	1

Table 2: Comparison of MXMap and RESIT-MLP on 10 randomly sampled segments from a 4-species chain system $w \Rightarrow x \Rightarrow y \Rightarrow z$.

To further assess prediction consistency, we apply the MXMap and RESIT (recapitulation in Appendix E) frameworks to different segments of the sequences to determine the causal order of variables. Specifically, we generate 10 random positive integers as starting timestamps to sample 10 test sequences from the 4-species chain system (Eq. **??**), each with a length of 3500.

For MXMap, we select a k-nearest neighbor (kNN) size of 10, a PCM correlation ratio threshold of 0.6, and delay embedding parameters $\tau = 2$ and $dim = 6$. For RESIT, we used the scikit-learn MLP regressor with two layers of 32 units each, along with an HSIC threshold ($\alpha = 0.01$) for edge removal.

Table 2 shows the results (evaluation metrics listed in Appendix H). MXMap consistently determined the correct causal order across all sampled segments, yielding stable predictions with perfect precision, recall, and F1 scores. In contrast, RESIT-MLP's predictions varied significantly depending on the starting point of each sequence. The causal graph predicted by RESIT changed across different segments, illustrating the sensitivity of its prediction to initial conditions due to the underlying predictive model assumption. Additionally, MXMap achieved better evaluation metrics across all four selected metrics.

4.4. Comparison with Other Established Causal Inference Methods

To demonstrate the effectiveness of MXMap in multivariate causal inference for nonlinear dynamical systems, we compare its performance on simulated multivariate dynamical systems (including cycles) with baseline methods (tsFCI, VAR-LiNGAM, PCMCI, Granger Causality, DYNOTEARS, SLARAC). When interpreting the predicted outputs, we consider non-oriented causal edges and bidirectional causal edges equivalent, which is in turn reflected in the metric calculation.

4.4.1. SIMULATED SYSTEMS

Structure	3V Chain		3V Immorality		3V No Cycle		3V Cycle	
	No Noise	Noise	No Noise	Noise	No Noise	Noise	No Noise	Noise
tsFCI	4	3	2	2	6	4	3	5
VARLiNGAM	2	4	4	4	5	6	3	4
Granger	4	4	4	4	5	6	2	4
PCMCI	**1**	**0**	**0**	**0**	2	1	3	3
DYNOTEARS	3	4	2	2	3	1	5	3
SLARAC	6	6	6	6	5	6	5	4
MXMap	**1**	**0**	1	**0**	**0**	**0**	**0**	**0**

Table 3: SHD scores of MXMap and baselines for 3V settings on simulated no-noise and noisy dynamical systems.

Tables 3, 4 and 5 show the results of SHD scores (best in bold) on simulated systems with varying complexity from 3 to 7 variables. A more complete evaluation with all four metrics ($Prec$, Rec, $F1$, and SHD), along with visualizations of ground truth graphs, predicted causal graphs, is provided in Appendix J. The time series data are generated under both noise-free and noisy settings (Gaussian additive noise, strength 0.01). Overall, MXMap consistently achieves good performance the baselines, yielding lower SHD scores which indicate fewer incorrect edges in the predicted causal graphs.

Structure	4V Chain		4V No Cycle		4V Cycle	
	No Noise	Noise	No Noise	Noise	No Noise	Noise
tsFCI	3	3	5	4	4	6
VARLiNGAM	4	6	7	8	4	5
Granger	4	6	5	8	3	6
PCMCI	1	**0**	**1**	**1**	8	4
DYNOTEARS	7	10	3	**1**	7	6
SLARAC	10	10	3	2	5	6
MXMap	**0**	**0**	**1**	**1**	**0**	2

Table 4: SHD scores of MXMap and baselines for 4V settings on simulated no-noise and noisy dynamical systems.

Structure	5V No Cycle		5V Cycle		6V No Cycle		7V Cycle	
	No Noise	Noise	No Noise	Noise	No Noise	Noise	No Noise	Noise
tsFCI	7	4	6	8	10	11	10	10
VARLiNGAM	6	6	12	12	9	11	16	15
Granger	6	6	12	12	11	12	15	15
PCMCI	5	5	5	**2**	11	**3**	11	10
DYNOTEARS	8	15	11	12	19	18	16	22
SLARAC	16	16	18	18	25	27	21	25
MXMap	**1**	**1**	**0**	**2**	**2**	4	**4**	**6**

Table 5: SHD scores of MXMap and baselines for 5V-7V settings on simulated no-noise and noisy dynamical systems.

4.4.2. ERA5 3-VARIABLE CHAIN: $tcw \Rightarrow rad \Rightarrow T_{2m}$
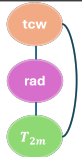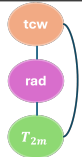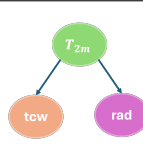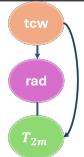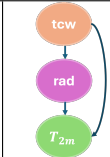
Method	PC	FCI	LiNGAM	PCMCI	MXMap
Output					
$Prec$	0.33	0.33	0	0.50	**0.67**
Rec	**1.0**	**1.0**	0	**1.0**	**1.0**
$F1$	0.50	0.50	0	0.67	**0.80**
SHD	4	4	4	2	**1**

Table 6: Causal inference methods on the ERA5 3V system.

For inferring the chain $tcw \Rightarrow rad \Rightarrow T_{2m}$, we take input sequence length of 6000, and consider the lag τ value to be 4 (for the lag value of delay embedding formulation, and max lag of the PCMCI method) While other methods either failed to identify causal directions correctly or predicted incorrect causal orders, MXMap consistently maintained the correct causal order and produced results closest to the expected ground truth.

Methods	tsFCI	VARLiNGAM	Granger	PCMCI	DYNOTEARS	SLARAC	MXMap
$rad_{solar} \Rightarrow T_{2m}$	✓	✓(half)	✓(half)	✓(half)	✓(half)	✗	✓
$rad_{terr} \Rightarrow T_{2m}$	✓(half)	✓(half)	✓	✓(half)	✗	✓	✓
$T_{adv950} \Rightarrow T_{2m}$	✗	✓(half)	✓(half)	✓(half)	✓	✓(half)	✓
$tcw \Rightarrow rad_{solar}$	✗	✗	✓(half)	✓	✓	✓	✓

Table 7: Detection of Benchmark Causal Relationships in the ERA5 5V System (full visualizations in Table 9). A checkmark (green) indicates a correctly detected and oriented edge, a half-checkmark (gray) denotes a detected but ambiguously oriented edge, and a crossmark (red) represents an undetected or incorrectly oriented edge.

4.4.3. ERA5 5-VARIABLE SYSTEM: tcw, rad_{solar}, rad_{terr}, T_{adv950} AND T_{2m}

The objective is to evaluate the performance via observing how many of these well-established causal relationships are detected by the methods $rad_{solar} \Rightarrow T_{2m}$, $rad_{terr} \Rightarrow T_{2m}$, $T_{adv950} \Rightarrow T_{2m}$, $tcw \Rightarrow rad_{solar}$ (explanations of these mechanisms in Appendix D, full table with predicted graphs in Table 9 in Appendix D.4). The results in Table 7 show that MXMap is overall able to correctly identify and orient edges that represent the 4 benchmark causal mechanisms in the 5-variable system, and outperforms the other baseline methods.

5. Conclusion

In this work, we proposed multiPCM, allowing us to more effectively distinguish between direct and indirect causal relationships. We integrated multiPCM with bivariate Convergent Cross Mapping (CCM) in a two-phase framework, MXMap, that first establishes an initial causal graph and then prunes indirect connections. Through experiments on simulated species interaction systems and real-world ERA5 meteorological data, we demonstrated that MXMap outperforms traditional methods and exhibits robust performance in complex, high-dimensional dynamical systems.

There are still limits to this current framework (discussed in Appendix I), demonstrated in runtime complexity, scalability, and possible failure cases in highly noisy environment or non-stationary systems. Future work could focus on enhancing the robustness of cross mapping under noisy conditions (Mønster et al., 2017). Investigating and incorporating certain noise-handling mechanisms (Zhang et al., 2024) in MXMap could further enhance its applicability in noisy real-world scenarios. Another direction is to explore adaptive parameter selection (Shortreed and Ertefaie, 2017; Machlanski et al., 2023) for MXMap, such as optimizing embedding dimensions and lags based on data properties and currently outputs. Current grid search methods are computationally expensive for larger datasets, and efficient heuristic or learning-based tuning could improve scalability. Finally, applying MXMap to other real-world domains, such as power systems, larger timescale climate modeling, and epidemiology, could further validate its versatility and reveal complex causal interactions.

Acknowledgments

References

Frédéric Barraquand, Coralie Picoche, Matteo Detto, and Florian Hartig. Inferring species interactions using granger causality and convergent cross mapping. *Theoretical Ecology*, 14(1):87–105, 2021.

Paul Beaumont, Ben Horsburgh, Philip Pilgerstorfer, Angel Droth, Richard Oentaryo, Steven Ler, Hiep Nguyen, Gabriel Azevedo Ferreira, Zain Patel, and Wesley Leong. CausalNex, October 2021. URL https://github.com/quantumblacklabs/causalnex.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

Kurt Butler, Guanchao Feng, and Petar M Djurić. On causal discovery with convergent cross mapping. *IEEE Transactions on Signal Processing*, 2023.

Gustau Camps-Valls, Andreas Gerhardus, Urmi Ninad, Gherardo Varando, Georg Martius, Emili Balaguer-Ballester, Ricardo Vinuesa, Emiliano Diaz, Laure Zanna, and Jakob Runge. Discovering causal relations and equations from data. *Physics Reports*, 1044:1–68, 2023.

Ziyue Chen, Miaoqing Xu, Bingbo Gao, George Sugihara, Feixue Shen, Yanyan Cai, Anqi Li, Qi Wu, Lin Yang, Qi Yao, et al. Causation inference in complicated atmospheric environment. *Environmental Pollution*, 303:119057, 2022.

Bree Cummins, Tomás Gedeon, and Kelly Spendlove. On the efficacy of state space reconstruction methods in determining causality. *SIAM Journal on Applied Dynamical Systems*, 14(1):335–381, 2015.

Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. *Probabilistic graphical models*, 16, 2010.

Andrew M Fraser and Harry L Swinney. Independent coordinates for strange attractors from mutual information. *Physical review A*, 33(2):1134, 1986.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(5), 2010.

Qingchao Jiang, Jiashi Jiang, Wenjing Wang, Chunjian Pan, and Weimin Zhong. Partial cross mapping based on sparse variable selection for direct fault root cause diagnosis for industrial processes. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Matthew B Kennel, Reggie Brown, and Henry DI Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A*, 45(6):3403, 1992.

Grigor Keropyan, David Strieder, and Mathias Drton. Rank-based causal discovery for post-nonlinear models. In *International Conference on Artificial Intelligence and Statistics*, pages 7849–7870. PMLR, 2023.

B Kliková and Aleš Raidl. Reconstruction of phase space of dynamical systems using method of time delay. In *Proceedings of WDS*, volume 11, pages 83–87, 2011.

Dimitris Kugiumtzis. State space reconstruction parameters in the analysis of chaotic time series—the role of the time window length. *Physica D: Nonlinear Phenomena*, 95(1):13–28, 1996.

Siyang Leng, Huanfei Ma, Jürgen Kurths, Ying-Cheng Lai, Wei Lin, Kazuyuki Aihara, and Luonan Chen. Partial cross mapping eliminates indirect causal influences. *Nature communications*, 11(1): 2632, 2020.

Wenqin Liu, Biwei Huang, Erdun Gao, Qiuhong Ke, Howard Bondell, and Mingming Gong. Causal discovery with mixed linear and nonlinear additive noise models: A scalable approach. In *Causal Learning and Reasoning*, pages 1237–1263. PMLR, 2024.

AJ Lotka. Elements of physical biology. *Williams and Wilkins*, 1925.

Damian Machlanski, Spyridon Samothrakis, and Paul Clarke. Hyperparameter tuning and model evaluation in causal effect estimation. *arXiv preprint arXiv:2303.01412*, 2023.

John Milnor. On the concept of attractor. *Communications in Mathematical Physics*, 99:177–195, 1985.

Dan Mønster, Riccardo Fusaroli, Kristian Tylén, Andreas Roepstorff, and Jacob F Sherson. Causal inference from noisy time-series data—testing the convergent cross-mapping algorithm in the presence of noise and external influence. *Future Generation Computer Systems*, 73:52–62, 2017.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. Pmlr, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Roger A Pielke Sr and Toshihisa Matsui. Should light wind and windy nights have the same temperature trends at individual levels even if the boundary layer averaged heat content change is the same? *Geophysical Research Letters*, 32(21), 2005.

VLRD Ramanathan, RD Cess, EF Harrison, P Minnis, BR Barkstrom, E Ahmad, and D Hartmann. Cloud-radiative forcing and climate: Results from the earth radiation budget experiment. *Science*, 243(4887):57–63, 1989.

Lionel Roques and Mickaël D Chekroun. Probing chaos and biodiversity in a simple competition model. *Ecological Complexity*, 8(1):98–104, 2011.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5 (11):eaau4996, 2019.

Jakob Runge, Xavier-Andoni Tibau, Matthias Bruhns, Jordi Muñoz-Marí, and Gustau Camps-Valls. The causality for climate competition. In *NeurIPS 2019 Competition and Demonstration Track*, pages 110–120. Pmlr, 2020.

Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505, 2023.

Tim Sauer, James A Yorke, and Martin Casdagli. J stat phys. *Embedology*, 65(3-4):579–616, 1991.

Shohei Shimizu. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

Peter L Spirtes, Christopher Meek, and Thomas S Richardson. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

Graeme L Stephens. Cloud feedbacks in the climate system: A critical review. *Journal of climate*, 18(2):237–273, 2005.

George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012.

Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*, pages 366–381. Springer, 2006.

I Vlachos and D Kugiumtzis. State space reconstruction for multivariate time series prediction. *arXiv preprint arXiv:0809.2220*, 2008.

Ioannis Vlachos and Dimitris Kugiumtzis. Nonuniform state-space reconstruction and coupling detection. *Physical Review E*, 82(1):016207, 2010.

Vito Volterra. *Théorie mathématique de la lutte pour la vie*. Gauthiers-Villars, 1931.

Sebastian Weichwald, Martin E Jakobsen, Phillip B Mogensen, Lasse Petersen, Nikolaj Thams, and Gherardo Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In *NeurIPS 2019 Competition and Demonstration Track*, pages 27–36. PMLR, 2020.

Hassler Whitney. Differentiable manifolds. *Annals of Mathematics*, pages 645–680, 1936.

Elise Zhang, François Mirallès, Raphaël Rousseau-Rizzi, Arnaud Zinflou, Benoit Boulet, and Di Wu. Enhancing convergent cross mapping: Simple preprocessing for noise-resilient causal discovery. *Authorea Preprints*, 2024.

Kun Zhang, Zhikun Wang, Jiji Zhang, and Bernhard Schölkopf. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024.

Appendix A. MXMap Workflow in Pseudocode

Algorithm 1 MXMap Workflow

Data: Delay τ, dimension E; Time series of a multivariate dynamical system $G = \{X_1, X_2, \ldots, X_n\}$; Variable indices $S := \{1, \ldots, K\}$; PCM threshold of correlation ratio $\gamma^\star$.

Result: Adjacency matrix $\mathcal{A}$; Final children dictionary CH

Phase 1: Initial Causal Graph

for $i \in S$ **do**

    for $j \in S$ *and* $j \neq i$ **do**

        Embed X_i and X_j with delay parameters τ and E; Perform cross mappings to reconstruct X_i from X_j and vice versa;

        Compute scores $\beta_{X_i \Rightarrow X_j}$ and $\beta_{X_j \Rightarrow X_i}$;

        if $\beta_{X_i \Rightarrow X_j} > \beta_{X_j \Rightarrow X_i}$ *and both scores* ≥ 0.5 **then**

            | Establish link $X_i \Rightarrow X_j$ (set $\mathcal{A}_{i,j} = 1$ and add j to $CH[i]$);

        end

        else if $\beta_{X_j \Rightarrow X_i} > \beta_{X_i \Rightarrow X_j}$ *and both scores* ≥ 0.5 **then**

            | Establish link $X_j \Rightarrow X_i$ (set $\mathcal{A}_{j,i} = 1$ and add i to $CH[j]$);

        end

    end

end

Phase 2: Prune Indirect Connections

for $i \in S$ **do**

    for $j \in CH[i]$ **do**

        if *Longest path between X_i and X_j has more than 2 nodes* **then**

            Identify all intermediate nodes between X_i and X_j as set *Conds*;

            Embed X_i, X_j, and *Conds*; perform multiPCM to compute reconstructions;

            Calculate scores ρ_{All}, ρ_{Direct}, and ratio $\gamma = \rho_{Direct}/\rho_{All}$;

            if $\gamma < \gamma^\star$ **then**

                | Remove link $X_i \Rightarrow X_j$ (set $\mathcal{A}_{i,j} = 0$ and remove j from $CH[i]$);

            end

        end

    end

end

Appendix B. Validation of multiPCM with More Generated Systems

A more complete grid-search results is presented in Table 8, with more cases of simulated systems and additional profiles of correlation scores ρ_{All} and ρ_{Direct}, on top of the ratio and label.

Input length is $L = 3500$, the ranges of lags and embedding dimensions are ($\tau, E \in \{1, 2, 3, \ldots, 8\}$). The predicted labels indicating whether direct causality is rejected based on a PCM threshold of 0.45 (this threshold selection is discussed in Appendix C). Red label indicates rejection, suggesting there is only indirect causality, while blue label indicates acceptance, suggesting direct causality exists hence we should not remove the link between the colored nodes.

18
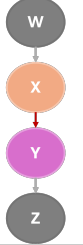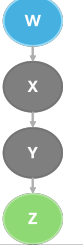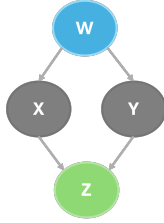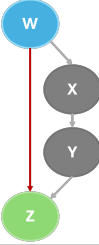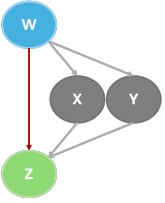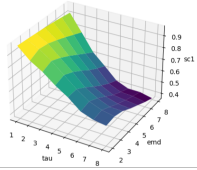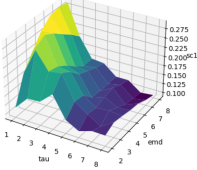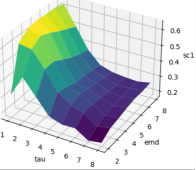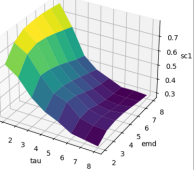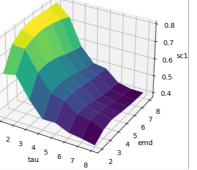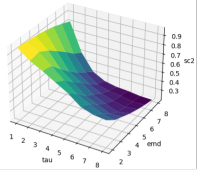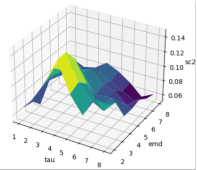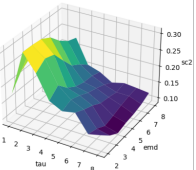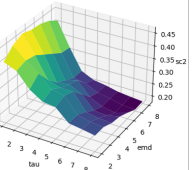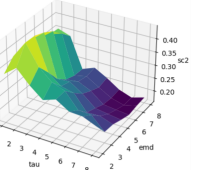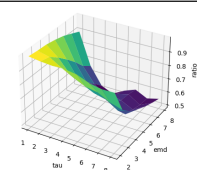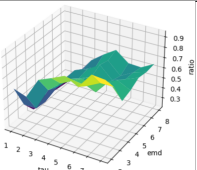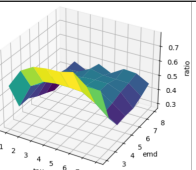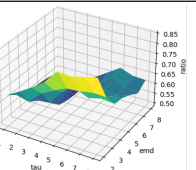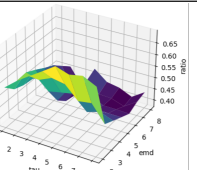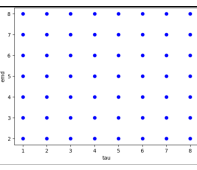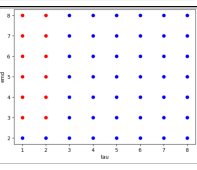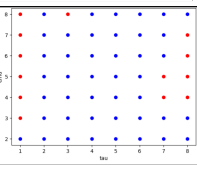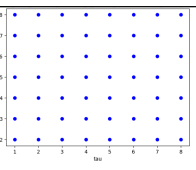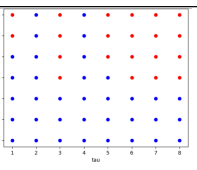
Table 8: Performance of multiPCM: Correlation scores, correlation ratio, predicted label ($thres =$ 0.45) under grid search. Red dot indicates there isn't direct causality between the colored nodes, while blue indicates there is direct causality.

Appendix C. Empirical Threshold Selection for multiPCM in Simulated Systems

The threshold for correlation ratio used in multiPCM was selected empirically by testing on our simulated systems under three distinct causality scenarios: direct, indirect, and both. The data of these scenarios are generated based on the description of Section 4.1.1, varying from 3-variable to 7-variable with increasing complexity. The data generation is noise-free for simplicity.

Here we use lag $\tau = 1$ and embedding dimension $E = 7$ for multiPCM (since the discrete data generation uses 1 as the lag, and the highest dimension of all the systems is 7, we round the dimension up to the highest value to ensure performance across all simulated systems). We illustrate the output labels with varying ratio thresholds ranging from 0.05 to 0.95, with a step size of 0.05. Lower ratio threshold implies that the edge removal is more lenient and tends to retain more edges

that lie between direct and indirect; while higher threshold indicates the edge removal is stricter and will only keep the edges whose direct correlation is strong enough.

For direct and both causality scenarios, the link should be retained (labeled "direct"); For indirect causality scenarios, the link should be removed (labeled "indirect").

The selected threshold should achieve a relatively good accuracy across all systems, here we set the tolerance for each threshold value to 2 mistaken labels at each causality senario.

(*a*) Direct (expected "direct") (*b*) Indirect (expected "indirect") (*c*) Both (expected "direct")
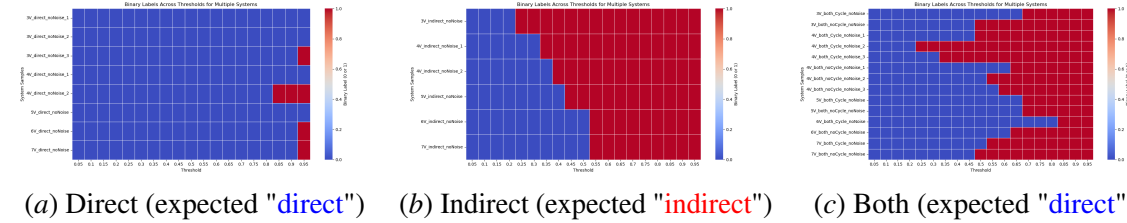
Figure 2: Different PCM ratio thresholds in three causality scenarios and how they impact the predicted labels.

Fig. 2 illustrates the results in heatmaps. In the "Direct" scenario, the thresholds that satisfy the tolerance are from 0.05 all the way to 0.9 (inclusive). In the "Indirect" scenario, the thresholds that satisfy the tolerance are from 0.45 to 0.95 (inclusive). In the "Both" scenario, the thresholds that satisfy the tolerance are from 0.05 to 0.45 (inclusive). Overall, 0.45 is relatively the best threshold value across all the simulation scenarios.

Note that the appropriate threshold may vary depending on the system and the specific purpose of the study. For the ERA5 chain $tcw \Rightarrow rad \Rightarrow T_{2m}$ discussed in Section 4.4.2, with MXMap lag $\tau = 4$ and dimension $E = 6$, the PCM ratio threshold required to remove the likely-indirect link $tcw \Rightarrow T_{2m}$ is 0.7. This higher threshold value may suggest the presence of a weaker but direct causal relationship between total cloud water tcw and near-ground temperature T_{2m}, or it could reflect the influence of unaccounted latent variables. If the goal is to establish a causal graph with the correct order while tolerating some additional links, a denser graph may be acceptable for certain applications. In such cases, tolerating the retention of certain edges by multiPCM can still yield useful insights.

Appendix D. ERA5 Meteorological Data: Candidate Testing Systems

D.1. Overview

This paper evaluates the causal discovery performance using the following two setups:

- **3V Chain**: A simplified 3-variable system capturing the causal chain: $tcw \Rightarrow rad \Rightarrow T_{2m}$. This causal chain is particularly pronounced in winter when cloud coverage strongly affects radiation levels, which subsequently modulate ground-level temperatures.

- **5V System**: This system includes solar radiation (rad_{solar}), terrestrial radiation (rad_{terr}), near-ground temperature advection (T_{adv950}), total cloud water (tcw) and near-ground temperature (T_{2m}). This setup examines whether well-established causal relationships can be detected, namely $rad_{solar} \Rightarrow T_{2m}$, $rad_{terr} \Rightarrow T_{2m}$, $T_{adv950} \Rightarrow T_{2m}$, $tcw \Rightarrow rad_{solar}$. The 5V system builds on the 3V chain by introducing two radiation components as well as a temperature

advection term. While the complete causal graph of this system is not fully established due to its complexity, the bivariate relationships outlined above are well-supported in meteorological literature and provide a robust benchmark for evaluation.

Detailed explanations are as follows.

D.2. Cloud (tcw) $\Rightarrow$ Radiation (rad) $\Rightarrow$ Ground-Level Temperature (T_{2m})

This causal chain used in our work, where clouds impact radiation levels and in turn affect ground-level temperatures, is well-supported in meteorological research, especially concerning winter seasons.

Clouds play a significant role in modulating surface radiation, as they both reflect incoming solar radiation (shortwave) and trap outgoing terrestrial radiation (longwave) (Stephens, 2005), hence $tcw \Rightarrow rad$.

During the day, clouds primarily act as a barrier to incoming solar radiation. Thick and extensive cloud cover reflects a significant portion of the solar radiation back into space, thereby lowering the temperature. At night, the longwave terrestrial radiation play a predominant role, which, combined with the presence of cloud cover (like an insulating blanket), leads to a warming effect.

The warming effect is especially pronounced in winter (Ramanathan et al., 1989), because the reduced daylight and lower solar angles reduce the overall shortwave radiation, rendering cloud cover's longwave effect more prominent. Research also indicates that during winter, the impact of clouds on radiation and temperature is more pronounced in polar regions, or in temperate zones with longer nights and lower solar angle, emphasizing the effect of cloud cover on maintaining warmer near-ground temperatures.

The timescale of this mechanism is usually within an hour or within a couple of hours (Stephens, 2005; Pielke Sr and Matsui, 2005): the impact of cloud cover on radiation flux is almost instantaneous, and the response of ground-level temperature depends on several factors (e.g. time of the day, humidity, wind). This information is crucial for choosing the lag value for the state-space reconstruction used in cross mapping based methods and the maximal lag τ used in PCMCI.

D.3. Introducing More Variables to the System

To enhance the complexity and realism of the ERA5 testing system, we introduce additional variables that reflect key meteorological processes. These include the following:

D.3.1. NEAR-GROUND TEMPERATURE ADVECTION T_{adv950}

T_{adv950} stands for the near-ground temperature advection at the $950hPa$ pressure level. It captures the transport of temperature influenced by atmospheric motion, and should be a direct cause of the near-ground temperature T_{2m}, hence $T_{adv950} \Rightarrow T_{2m}$.

D.3.2. RADIATION COMPONENTS: SOLAR RADIATION (rad_{solar}) AND TERRESTRIAL RADIATION (rad_{terr})

Separating the radiation components into solar and terrestrial radiation enables finer analysis of their individual roles in energy balance and temperature dynamics:

- **Solar Radiation** (rad_{solar}) is mainly and directly affected by cloud cover due to its role in shielding or transmitting shortwave radiation, hence $tcw \Rightarrow rad_{solar}$.

- **Terrestrial Radiation** (rad_{terr}) and cloud cover have a more complex interaction mechanism, and their causal relationship would be weaker and less prominent compared to $tcw \Rightarrow rad_{solar}$.

Both radiation components should be directly impacting the temperature, hence $rad_{solar} \Rightarrow T_{2m}$ and $rad_{terr} \Rightarrow T_{2m}$.

D.4. Evaluating Causal Relationships

We evaluate the model performance on the following systems:

1. **The 3-variable chain** $tcw \Rightarrow rad \Rightarrow T_{2m}$**:** Since the ground truth is known and relatively well justified, the causal discovery of this system will be evaluated in the same way as the simulated systems, using the metrics to compute the differences between truth causal graph and predicted graph;

2. **The expanded 5-variable system with** tcw**,** rad_{solar}**,** rad_{terr}**,** T_{adv950} **and** T_{2m}**:** While establishing the exact ground truth causal graph for this higher-dimensional weather system is challenging due to its inherent complexity, the following bivariate causal relationships are well-understood and serve as a benchmark for evaluation: $rad_{solar} \Rightarrow T_{2m}$, $rad_{terr} \Rightarrow T_{2m}$, $T_{adv950} \Rightarrow T_{2m}$, $tcw \Rightarrow rad_{solar}$; We evaluate how well the causal inference methods detect these known relationships consistently.

We present the full results along with the predicted graphs in Appendix Table 9 to supplement the results presented in main Section 4.4.3.

Appendix E. RESIT (Regression with Subsequent Independence Test) Framework

RESIT is an approach that extends the principles of ANM by iteratively conducting regression followed by an independence test on the residuals (Peters et al., 2014). The original RESIT framework contains two phases, the causal order determination phase, and the edge elimination phase and is outlined in Algorithm 2.

In Phase 1, one variable is selected as a prediction target (alleged effect) in each iteration, and the remaining variables are used as predictors (alleged causes) to fit a regression model. The regression error is computed, and the dependence between the residue and the predictors is measured. The variable with the weakest dependence (most likely effect) is identified and removed from the set. This process is repeated until all variables are ordered, establishing a topological order.

In Phase 2, for each variable with at least one parent, the set of parent variables is retrieved. For each parent variable, a regression model is fitted to predict the target variable using all other parents except the current one. The independence of the regression residue is compared with a threshold, and if it is sufficiently independent, the parent variable is removed from the parent set. This phase refines the causal structure by eliminating unnecessary dependencies, resulting in a clearer causal graph.

The RESIT algorithm iteratively establishes a causal order of variables by evaluating the strength of their causal relationships through regression and independence testing, ultimately refining the identified causal structure.

Table 9: Full results for detection of Benchmark Causal Relationships in the ERA5 5V System. A checkmark (green) indicates a correctly detected and oriented edge, a half-checkmark (gray) denotes a detected but ambiguously oriented edge, and a crossmark (red) represents an undetected or incorrectly oriented edge.

Methods	tsFCI	VARLiNGAM	Granger	PCMCI
Output				
$rad_{solar} \Rightarrow T_{2m}$	✓	✔	✔	✔
$rad_{terr} \Rightarrow T_{2m}$	✔	✔	✓	✔
$T_{adv950} \Rightarrow T_{2m}$	✗	✔	✔	✔
$tcw \Rightarrow rad_{solar}$	✗	✗	✔	✓

Methods	DYNOTEARS	SLARAC	MXMap
Output			
$rad_{solar} \Rightarrow T_{2m}$	✔	✗	✓
$rad_{terr} \Rightarrow T_{2m}$	✗	✓	✓
$T_{adv950} \Rightarrow T_{2m}$	✓	✔	✓
$tcw \Rightarrow rad_{solar}$	✓	✓	✓

Appendix F. Overview of the Baseline Methods

Time-Series Fast Causal Inference (tsFCI) (Entner and Hoyer, 2010) is an extension of the Fast Causal Inference (FCI) (Spirtes et al., 2013) algorithm for time-series data. It is designed to infer causal relationships from temporal data, taking into account potential latent confounders. The output is a Partial Ancestral Graph (PAG) with repeating structures over time steps to reflect the time-lagged causal impact. The predicted result allows unoriented edges, here we interpret such edge as bidirectional.

VAR-LiNGAM (Vector Autoregressive Linear Non-Gaussian Acyclic Model) (Hyvärinen et al., 2010) enhances the LiNGAM (Shimizu et al., 2006) framework with vector autoregressive (VAR) models, enabling causal discovery in multivariate time series assuming non-Gaussian distribution and acyclic causal graph.

PCMCI (Peter and Clark Momentary Conditional Independence) (Runge et al., 2019), is a method designed to detect causal links in time series data. PCMCI extends the PC algorithm for temporal data, combining it with momentary conditional independence tests to account for lagged dependencies and reduce the risk of false discoveries in highly autocorrelated data. PCMCI identifies causal links at different levels of time lags, hence its time-unrolled causal graphs will be fully oriented, while the

23

Algorithm 2 Regression with Subsequent Independence Test (RESIT)

Input: i.i.d. samples of a p-dimensional distribution on $(X_1, \ldots, X_p)$
Output: Set of parents for each variable: $\{\mathrm{pa}(1), \ldots, \mathrm{pa}(p)\}$
$S \leftarrow \{1, \ldots, p\}$
$\pi \leftarrow [\,]$
Phase 1: Determine topological order
while $S \neq \emptyset$ **do**
    for $k \in S$ **do**
        Regress X_k on $\{X_i\}_{i \in S \setminus \{k\}}$;
        Measure dependence between residual $r = X_k - \hat{X}_k$ and predictors $\{X_i\}_{i \in S \setminus \{k\}}$;
    end
    Let k^* be the variable with the weakest dependence ;
    $S \leftarrow S \setminus \{k^*\}$;
    $\mathrm{pa}(k^*) \leftarrow S$;
    $\pi \leftarrow [k^*, \pi]$;
end
Phase 2: Remove superfluous edges
for $k = 2$ **to** p **do**
    for $\ell \in pa(\pi(k))$ **do**
        Regress $X_{\pi(k)}$ on $\{X_i\}_{i \in \mathrm{pa}(\pi(k)) \setminus \{\ell\}}$;
        if *residuals are independent of* $\{X_i\}_{i \in \{\pi(1), \ldots, \pi(k-1)\}}$ **then**
           $\mathrm{pa}(\pi(k)) \leftarrow \mathrm{pa}(\pi(k)) \setminus \{\ell\}$;
        end
    end
end

---

conventional causal graph may have contemporaneous, bidirectional or feedback processes. In our work, such link is interpreted as bidirectional.

Granger Causality (Granger, 1969) is a statistical hypothesis test used to determine whether one time-series can predict another, implying a causal relationship. It is based on the idea that a time-series $\mathbf{X}$ is said to Granger-cause another series $\mathbf{Y}$ if including past values of $\mathbf{X}$ improves the prediction of $\mathbf{Y}$ over a model that only uses past values of $\mathbf{Y}$.

DYNOTEARS (Pamfil et al., 2020) is a Bayesian structure learning algorithm for time series data. It extends the DAG-NOTEARS (Zheng et al., 2018) framework, allowing for detection of contemporaneous and time-lagged relationships between variables in a time-series. The method enforces acyclicity via a differentiable constraint, enabling scalable and accurate causal discovery.

SLARAC (Subsampled Linear Auto-Regression Absolute Coefficients) (Weichwald et al., 2020) is among the `tidybench` time-series causal discovery benchmarks. It is among the best performing models in the NeurIPS 2019 Causality for Climate competition (Runge et al., 2020), emphasizes the use of large regression coefficients over traditional small p-value thresholds.

Causal discovery with observational data is a rich and fast-developing field, offering diverse methods and baselines. For a comprehensive review of recent advances, please refer to (Camps-Valls et al., 2023), where methodologies, challenges and applications are discussed in depth.

Appendix G. Experimental Setup

RESIT is implemented with the `LiNGAM` library and MLP regressor from `scikit-learn` (Pedregosa et al., 2011; Buitinck et al., 2013). To ensure a good enough fit quality, we choose the following parameters for they MLP regressor: 2 hidden layers of size 100, max number of iteration 1000.

tsFCI is implemented with FCI from `causal-learn` using the `fisherz` as independence test and level of significance $\alpha = 0.05$; VAR-LiNGAM and Granger implemented with `causal-learn` with max lag of 1; DYNOTEARS is implemented using the `Causalnex` library, and SLARAC from `tidybench` with max lag of 1. PCMCI from the `tigramite` package is implemented using robust correlation (`RobustCorr`), with a significance level of $\alpha_{pc} = 0.2$ for PC and a maximum lag of 1. As recommended by the original authors (Runge et al., 2019), α_{pc} should not be interpreted as a strict measure of statistical significance. Instead, it is used in Phase 1 (PC) to reduce the search space for independence tests, and a slightly larger value for α_{pc} often provides faster inference while maintaining relatively good accuracy.

Appendix H. Evaluation Metrics for Graph Outputs

The adjacency matrix of each output graph is defined with entries of 0 and 1, where the row index represents the cause variable and the column index represents the effect variable. An entry of 1 at $(row, column)$ indicates the existence of a causal edge between the two variables. To evaluate the model performance, the following metric scores are calculated on the adjacency matrices of the ground truth and the predicted graphs [2] :

- **Precision:** Precision measures the proportion of correctly identified causal edges out of all edges predicted by the model. Mathematically, it is defined as the number of true positive edges divided by the total number of predicted edges. In the context of adjacency matrices, it is calculated as:

$$\text{Precision} = \frac{\sum_{i,j} \mathbb{I}\left(\hat{A}_{ij} = 1 \text{ and } A_{ij} = 1\right)}{\sum_{i,j} \mathbb{I}\left(\hat{A}_{ij} = 1\right)} \tag{9}$$

  where $\hat{A}$ is the predicted adjacency matrix, A is the ground truth adjacency matrix, and $\mathbb{I}$ is the indicator function.

- **Recall:** Recall measures the proportion of correctly identified causal edges out of all actual edges in the ground truth. It is defined as the number of true positive edges divided by the total number of true edges. For adjacency matrices, recall is calculated as:

$$\text{Recall} = \frac{\sum_{i,j} \mathbb{I}\left(\hat{A}_{ij} = 1 \text{ and } A_{ij} = 1\right)}{\sum_{i,j} \mathbb{I}\left(A_{ij} = 1\right)} \tag{10}$$

- **F1:** The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. It is particularly useful when the class distribution is imbalanced. The

---

2. A tool for manual evaluation with a graphical interface has been implemented and can be found at https://github.com/elisejiuqizhang/DAGComparisonToolbox.

F1 score for adjacency matrices is calculated as:

$$\text{F1 Score } = 2 \times \frac{\text{Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}} \qquad (11)$$

- **Structural Hamming Distance (SHD):** It counts the number of edge additions, deletions, and reversals needed to transform the predicted adjacency matrix into the ground truth adjacency matrix. A lower SHD indicates a closer match to the true causal structure. Formally, SHD is calculated as:

$$\text{SHD} = \sum_{i,j} \mathbb{I}\left(\hat{A}_{ij} \neq A_{ij}\right) \qquad (12)$$

These four metrics provide a comprehensive evaluation of a model's performance in inferring causal structures from multivariate dynamical systems.

Appendix I. MXMap: Current Limits

I.1. Runtime Complexity

MXMap has a runtime complexity of $\mathcal{O}(n^2)$ due to its pairwise processing design. To validate the runtime behavior, we conduct additional experiments using the chain causal structure with an increasing number of variables (from 3V chain to 8V chain) and plot the runtime in Fig. 3, where we can confirm the $\mathcal{O}(n^2)$ complexity.
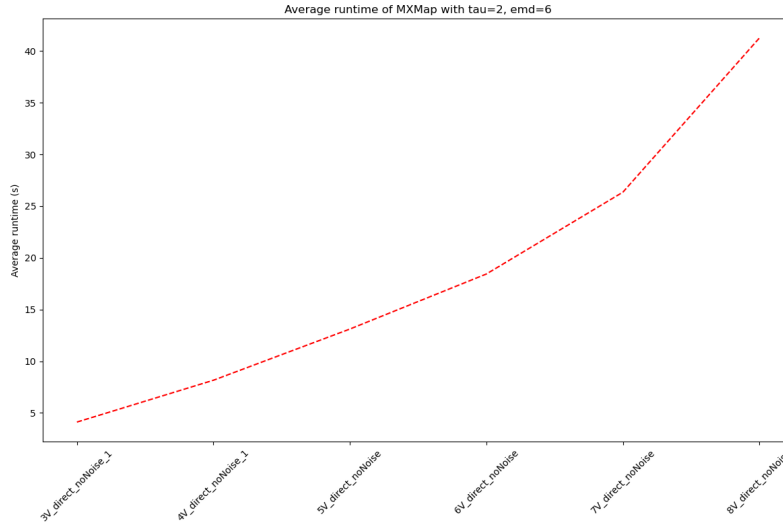
Figure 3: MXMap runtime as number of variables increase (chain structure).

To compare MXMap with the selected baseline methods, we track their runtime on the 8-variable chain (the highest dimensional system used in this work) with input length 4000. The approximate average runtimes were as follows:

- **MXMap:** 40 seconds

- **tsFCI:** 15 seconds

- **Granger:** 3 seconds

- **VAR-LiNGAM:** 4 seconds

- **DYNOTEARS:** 5 seconds

- **SLARAC:** 15 seconds

Although MXMap is not the fastest method, its causal detection performance remains the most satisfactory, since its assumptions are tailored to nonlinear coupled dynamical systems such as the simulated systems used in the experiments. The $\mathcal{O}(n^2)$ complexity may have limited the method's scalability, it is precisely this design that enables MXMap to reliably detect causal cycles, a capability lacking in many alternative methods. This can be interpreted as a trade-off between computational cost and performance highlights MXMap's ability to handle cyclic causal structures and achieve strong causal discovery in systems that match its assumptions.

I.2. Potential Failure Cases

MXMap, like other cross-mapping-based methods, relies on specific assumptions that may limit its applicability under certain conditions. The following outlines the potential failure cases for MXMap:

1. **System doesn't have an attractor:** The presence of an attractor is fundamental to all cross-mapping-based methods, ensuring that delay embedding reconstruction accurately captures the system's state-space structure. If such assumption is violated (e.g., transient systems, non-stationary time series, or systems dominated by stochasticity), there won't be reliable stable state-space structure to correctly establish time-index correspondence for nearest neighborhoods.

2. **Strong synchrony or forcing:** As discussed in the original CCM paper (Sugihara et al., 2012), strong forcing or synchrony can obscure the intrinsic dynamics of the forced (effect) variable, driving it to adopt the almost same manifold topology as the forcing (causal) variable. This compromises the reconstruction of independent state spaces, leading to unreliable causal inference.

3. **Improper lag τ and dimension E when constructing delay embeddings:** Incorrect choices of the delay-embedding parameters can hinder state-space reconstruction and lead to failure in CCM, as discussed by the tutorial with examples.

4. **High noise levels:** Observations with high noise may lead to failure in cross maps (Mønster et al., 2017).

Appendix J. Complete Causal Discovery Evaluation of MXMap and Baseline Methods on Simulated Systems

In this section, we present the complete analysis of performance for each model. The ground truth causal graphs and predicted graphs are visualized for each coupled system, and the four metrics are calculated for evaluation (best score in bold). In most of the cases, MXMap demonstrates optimal or second-best performance in discovering the underlying causal structure.
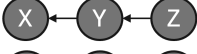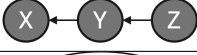
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0	0	0	4
VARLiNGAM			0.50	**1.0**	0.67	2
Granger			0	0	0	4
PCMCI			**0.67**	**1.0**	**0.80**	**1**
DYNOTEARS			0.40	**1.0**	0.57	3
SLARAC			0	0	0	6
MXMap			**0.67**	**1.0**	**0.80**	**1**

Table 10: 3V Chain (No Noise)
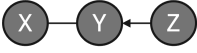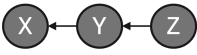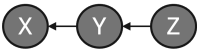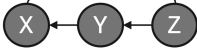
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.33	0.50	0.40	3
VARLiNGAM			0	0	0	4
Granger			0	0	0	4
PCMCI			**1.0**	**1.0**	**1.0**	**0**
DYNOTEARS			0.25	0.50	0.33	4
SLARAC			0	0	0	6
MXMap			**1.0**	**1.0**	**1.0**	**0**

Table 11: 3V Chain (Gaussian Additive Noise, Level 0.01)
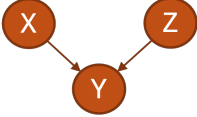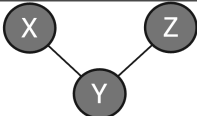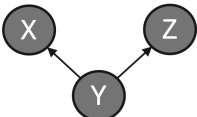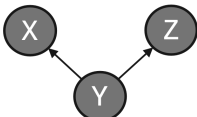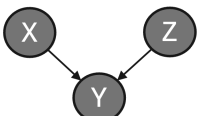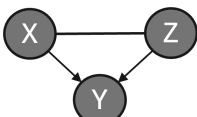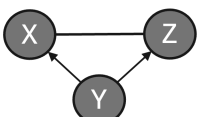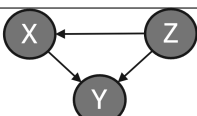
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.50	1.0	0.67	2
VARLiNGAM			0	0	0	4
Granger			0	0	0	4
PCMCI			**1.0**	**1.0**	**1.0**	**0**
DYNOTEARS			0.50	**1.0**	0.67	2
SLARAC			0	0	0	6
MXMap			0.67	**1.0**	0.80	**1**

Table 12: 3V Immorality (No Noise)
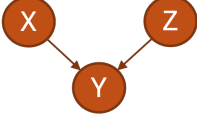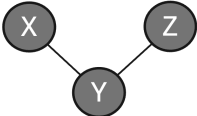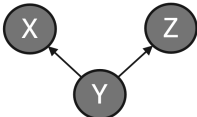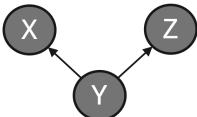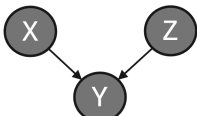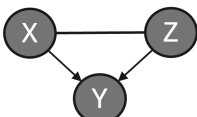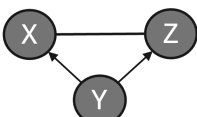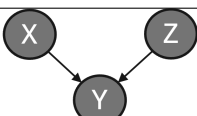
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.50	**1.0**	0.67	2
VARLiNGAM			0	0	0	4
Granger			0	0	0	4
PCMCI			**1.0**	**1.0**	**1.0**	**0**
DYNOTEARS			0.50	**1.0**	0.67	2
SLARAC			0	0	0	6
MXMap			**1.0**	**1.0**	**1.0**	**0**

Table 13: 3V Immorality (Gaussian Additive Noise, Level 0.01)
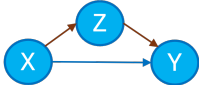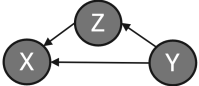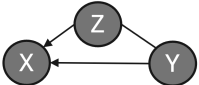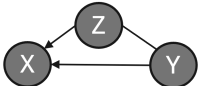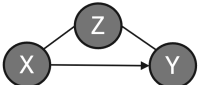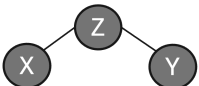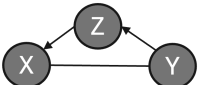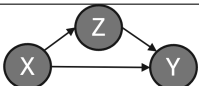
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0	0	0	6
VARLiNGAM			0.25	0.33	0.29	5
Granger			0.25	0.33	0.29	5
PCMCI			0.60	**1.0**	0.75	2
DYNOTEARS			0.50	0.67	0.57	3
SLARAC			0.25	0.33	0.29	5
MXMap			**1.0**	**1.0**	**1.0**	**0**

Table 14: 3V No Cycle (No Noise)
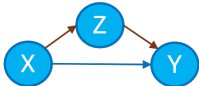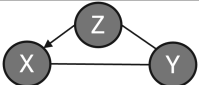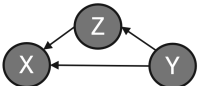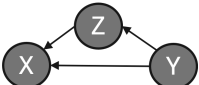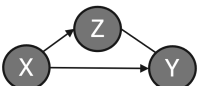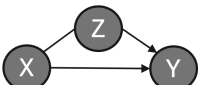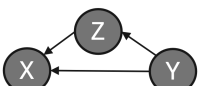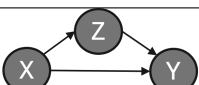
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.40	0.67	0.5	4
VARLiNGAM			0	0	0	6
Granger			0	0	0	6
PCMCI			0.75	**1.0**	0.86	1
DYNOTEARS			0.75	**1.0**	0.86	1
SLARAC			0	0	0	6
MXMap			**1.0**	**1.0**	**1.0**	**0**

Table 15: 3V No Cycle (Gaussian Additive Noise, Level 0.01)
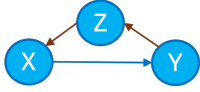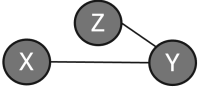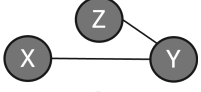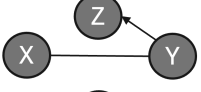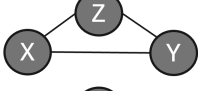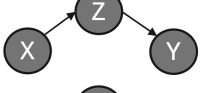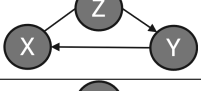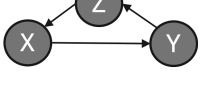
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.50	0.67	0.57	3
VARLiNGAM			0.50	0.67	0.57	3
Granger			0.67	0.67	0.67	2
PCMCI			0.50	**1.0**	0.67	3
DYNOTEARS			0	0	0	5
SLARAC			0.25	0.33	0.29	5
MXMap			**1.0**	**1.0**	**1.0**	**0**

Table 16: 3V Cycle (No Noise)
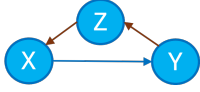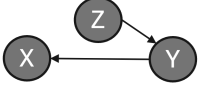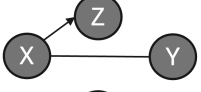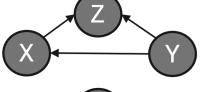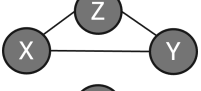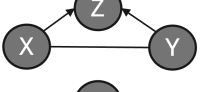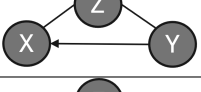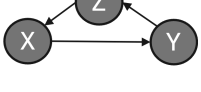
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0	0	0	5
VARLiNGAM			0.33	0.33	0.33	4
Granger			0.33	0.33	0.33	4
PCMCI			0.50	**1.0**	0.67	3
DYNOTEARS			0.50	0.65	0.57	3
SLARAC			0.40	0.67	0.50	4
MXMap			**1.0**	**1.0**	**1.0**	**0**

Table 17: 3V Cycle (Gaussian Additive Noise, Level 0.01)
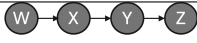
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.50	**1.0**	0.67	3
VARLiNGAM			0.40	0.67	0.50	4
Granger			0.33	0.33	0.33	4
PCMCI			0.75	**1.0**	0.86	1
DYNOTEARS			0.25	0.67	0.36	7
SLARAC			0.11	0.33	0.17	10
MXMap			**1.0**	**1.0**	**1.0**	**0**

Table 18: 4V Chain (No Noise)
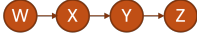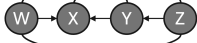
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.50	0.67	0.57	3
VARLiNGAM			0	0	0	6
Granger			0	0	0	6
PCMCI			**1.0**	**1.0**	**1.0**	**0**
DYNOTEARS			0.11	0.33	0.17	10
SLARAC			0.11	0.33	0.17	10
MXMap			**1.0**	**1.0**	**1.0**	**0**

Table 19: 4V Chain (Gaussian Additive Noise, Level 0.01)

Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.40	0.50	0.44	5
VARLiNGAM			0	0	0	7
Granger			0.40	0.50	0.44	5
PCMCI			0.80	**1.0**	**0.89**	**1**
DYNOTEARS			0.60	0.75	0.67	3
SLARAC			0.57	**1.0**	0.72	3
MXMap			**0.80**	**1.0**	**0.89**	**1**

Table 20: 4V No Cycle (No Noise)
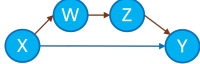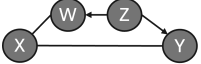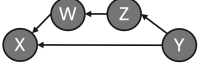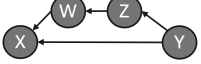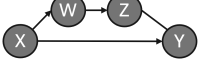
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.50	0.75	0.60	4
VARLiNGAM			0	0	0	8
Granger			0	0	0	8
PCMCI			**0.80**	**1.0**	**0.89**	**1**
DYNOTEARS			**0.80**	**1.0**	**0.89**	**1**
SLARAC			0.67	**1.0**	0.8	2
MXMap			**0.80**	**1.0**	**0.89**	**1**

Table 21: 4V No Cycle (Gaussian Additive Noise, Level 0.01)
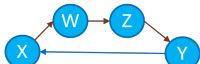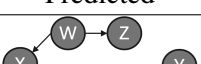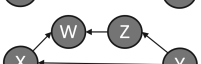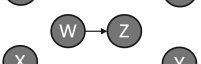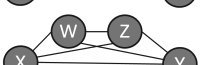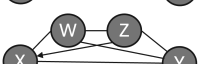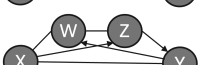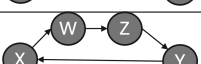
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.50	0.25	0.33	4
VARLiNGAM			0.50	0.50	0.50	4
Granger			**1.0**	0.25	0.40	3
PCMCI			0.33	**1.0**	0.50	8
DYNOTEARS			0.36	**1.0**	0.53	7
SLARAC			0.44	**1.0**	0.62	5
MXMap			**1.0**	**1.0**	**1.0**	**0**

Table 22: 4V Cycle (No Noise)

Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.25	0.25	0.25	6
VARLiNGAM			0.40	0.50	0.44	5
Granger			0.25	0.25	0.25	6
PCMCI			0.50	**1.0**	0.67	4
DYNOTEARS			0.38	0.75	0.50	6
SLARAC			0.38	0.75	0.50	6
MXMap			**0.67**	**1.0**	**0.80**	**2**

Table 23: 4V Cycle (Gaussian Additive Noise, Level 0.01)

Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.29	0.50	0.36	7
VARLiNGAM			0.25	0.25	0.25	6
Granger			0.25	0.25	0.25	6
PCMCI			0.42	0.75	0.55	5
DYNOTEARS			0.25	0.50	0.33	8
SLARAC			0.13	0.50	0.20	16
MXMap			**0.80**	**1.0**	**0.89**	**1**

Table 24: 5V Structure 1 Without Cycle (No Noise)
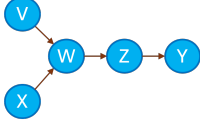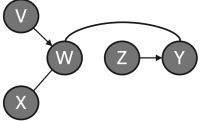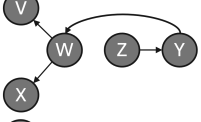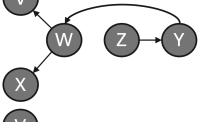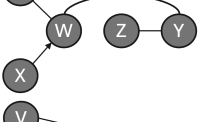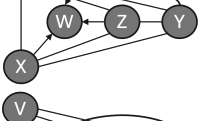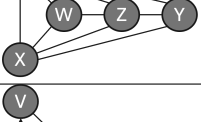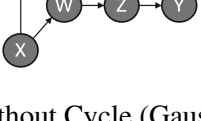
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.50	0.75	0.60	4
VARLiNGAM			0.25	0.25	0.25	6
Granger			0.25	0.25	0.25	6
PCMCI			0.43	0.75	0.55	5
DYNOTEARS			0.13	0.50	0.21	15
SLARAC			0.17	0.75	0.27	16
MXMap			**0.80**	**1.0**	**0.89**	**1**

Table 25: 5V Structure 1 Without Cycle (Gaussian Additive Noise, Level 0.01)
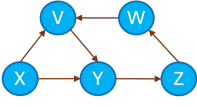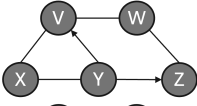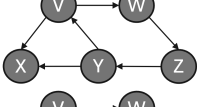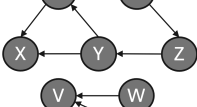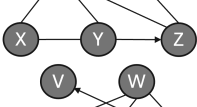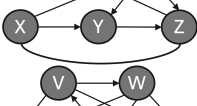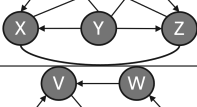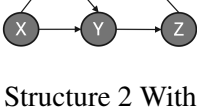
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.50	0.83	0.64	6
VARLiNGAM			0	0	0	12
Granger			0	0	0	12
PCMCI			0.55	**1.0**	0.71	5
DYNOTEARS			0.22	0.33	0.27	11
SLARAC			0.07	0.17	0.10	18
MXMap			**1.0**	**1.0**	**1.0**	**0**

Table 26: 5V Structure 2 With Cycle (No Noise)
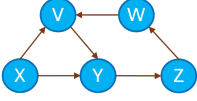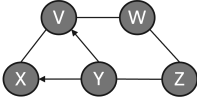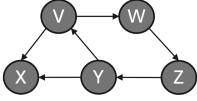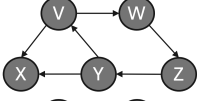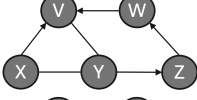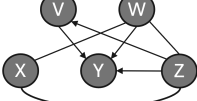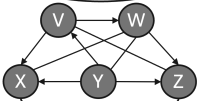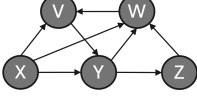
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.40	0.67	0.50	8
VARLiNGAM			0	0	0	12
Granger			0	0	0	12
PCMCI			**0.75**	**1.0**	**0.86**	**2**
DYNOTEARS			0.20	0.33	0.25	12
SLARAC			0.07	0.17	0.10	18
MXMap			**0.75**	**1.0**	**0.86**	**2**

Table 27: 5V Structure 2 With Cycle (Gaussian Additive Noise, Level 0.01)
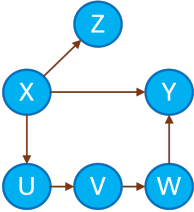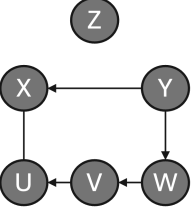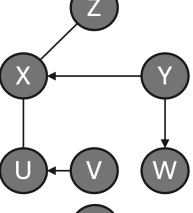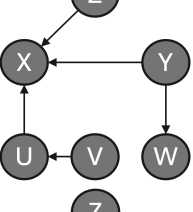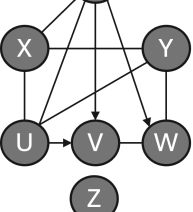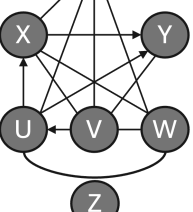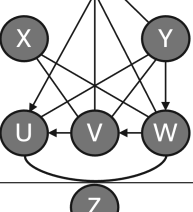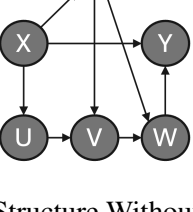
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.17	0.17	0.17	10
VARLiNGAM			0.29	0.33	0.31	9
Granger			0	0	0	11
PCMCI			0.35	**1.0**	0.52	11
DYNOTEARS			0.16	0.50	0.24	19
SLARAC			0	0	0	25
MXMap			**0.75**	**1.0**	**0.86**	**2**

Table 28: 6V Structure Without Cycle (No Noise)
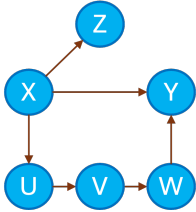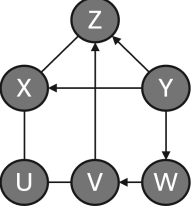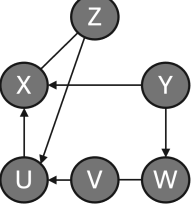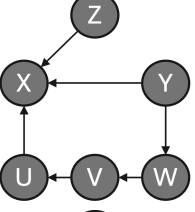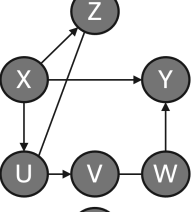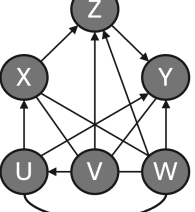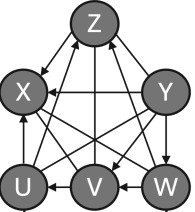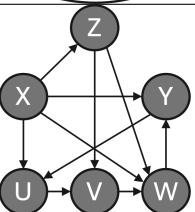
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.27	0.50	0.35	11
VARLiNGAM			0.22	0.33	0.27	11
Granger			0	0	0	12
PCMCI			**0.67**	**1.0**	**0.80**	**3**
DYNOTEARS			0.17	0.50	0.25	18
SLARAC			0	0	0	27
MXMap			0.60	**1.0**	0.75	4

Table 29: 6V Structure Without Cycle (Gaussian Additive Noise, Level 0.01)
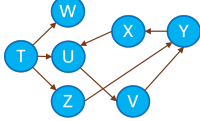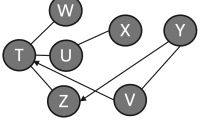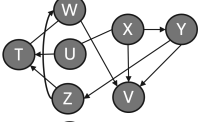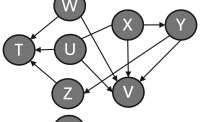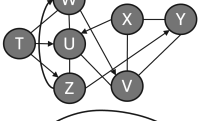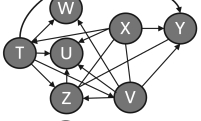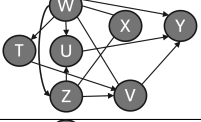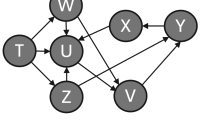
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.43	0.63	0.50	10
VARLiNGAM			0.17	0.25	0.20	16
Granger			0.18	0.25	0.21	15
PCMCI			0.42	**1.0**	0.59	11
DYNOTEARS			0.30	0.75	0.43	16
SLARAC			0.06	0.13	0.09	21
MXMap			**0.67**	**1.0**	**0.80**	**4**

Table 30: 7V Structure With Cycle (No Noise)
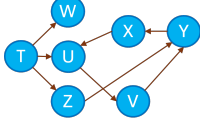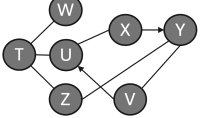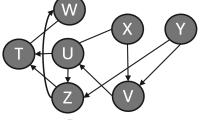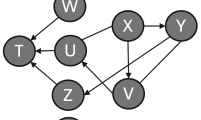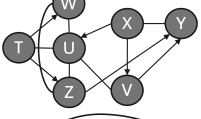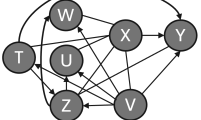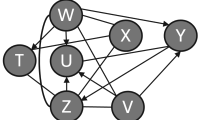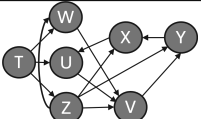
Method	Ground Truth	Predicted	Precision	Recall	F1	SHD
tsFCI			0.43	0.75	0.55	10
VARLiNGAM			0.18	0.25	0.21	15
Granger			0.18	0.25	0.21	15
PCMCI			0.44	**1.0**	0.62	10
DYNOTEARS			0.18	0.50	0.27	22
SLARAC			0.10	0.25	0.14	25
MXMap			**0.58**	0.88	**0.70**	**6**

Table 31: 7V Structure With Cycle (Gaussian Additive Noise, Level 0.01)