Learning General Causal Structures with Hidden Dynamic Process for Climate Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

The heart of climate analysis is a rational effort to understand the *causes* behind the *purely observational* data. Latent driving forces, such as atmospheric processes, play a critical role in temporal dynamics, and the task of inferring such latent forces is often a problem of Causal Representation Learning (CRL). Moreover, geographically nearby regions may directly interact with each other, and such direct causal relations among the observed data are often not modeled in traditional CRL, making the problem more challenging. In this paper, we propose a unified framework that can uncover not only the latent driving forces, but also the causal relations among the observed variables. We establish conditions under which the hidden dynamic process and the relations among the observed variables are simultaneously identifiable from time-series data. Even without parametric assumptions on the causal relations, we provide identifiability guarantees for recovering latent variables and the relations among the observed variables via contextual information. Guided by these insights, we propose a framework for nonparametric **Ca**usal **D**iscovery and **Re**presentation learning (**CaDRe**), based on a time-series generative model with structural constraints. Synthetic data validates our theoretical claims. On real-world climate datasets, CaDRe achieves competitive forecasting performance and offers the visualized causal graphs consistent with domain knowledge, which is expected to improve our understanding of the climate systems.

1 Introduction

Understanding the causal structure of climate systems is fundamental not only to scientific reasoning [68], but also to reliable modeling and prediction. Given the observed data with d_x variables: $\mathbf{x}_t = [x_{t,1}, \dots, x_{t,d_x}]$, our goal is twofold: (1) to discover the underlying latent variables $\mathbf{z}_t = [z_{t,1}, \dots, z_{t,d_z}]$ and their temporal interactions, and (2) to identify causal relations among observed variables. To better understand this problem, we describe it using a causal modeling perspective. As depicted in Figure 1, latent drivers $\mathbf{z}_t$, such as pressure and precipitation [9], are not directly measured but significantly influence the observed dynamics. These latent processes evolve jointly and stochastically, exhibiting both *instantaneous* and *time-lagged* causal dependencies [51, 66]. They govern observable quantities $\mathbf{x}_t$ like temperature, which reflect underlying dynamics and also exhibit spatial interactions through emergent weather patterns, such as wind circulation systems.

Identifying these underlying hidden variables and temporal relations is the central objective of Causal Representation Learning (CRL) [71] problem. Recent advances in identifiability theory and practical algorithm design fall under the framework of nonlinear Independent Component Analysis (ICA). These approaches typically rely on auxiliary variables [28, 29, 27, 86], sparsity [39, 97, 98, 38, 6], or restricted generative functions [18], and generally assume a *noise-free* and *invertible* generation from $\mathbf{z}_t$ to $\mathbf{x}_t$, in order to *directly* recover latent space. However, climatic measurements exhibit

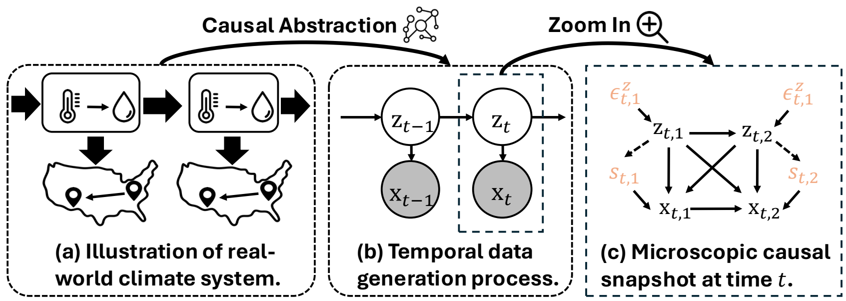


Figure 1: From climate system to causal graph. $\mathbf{x}_t$ represent observed data and $\mathbf{z}_t$ denotes unobserved variables behind $\mathbf{x}_t$, ϵ_t^z denotes the stochasticity in latent causal process, and s_t denotes the noise variable varying with $\mathbf{z}_t$, *e.g.*, human activities [9].

both observational dependencies and stochastic noise, violating these assumptions and limiting the applicability of existing CRL approaches.

This problem can also be cast as the problem of causal discovery [75, 59] in the presence of latent processes. Causal discovery often relies on parametric models, such as linear non-Gaussianity [72], nonlinear additive [23, 37], post-nonlinear models [92], as well as nonparametric methods with [26, 64, 54] or without auxiliary variables [74, 96, 93]. However, generally speaking, they cannot identify latent variables, their interrelations, and their causal influence on observed variables. For example, Fast Causal Inference (FCI) algorithm [74] produces asymptotically correct results in the presence of latent confounders by exploiting conditional independence relations, but its result is often not informative enough; for instance, it cannot recover causally-related latent variables.

This above underscores the need for a unified framework capable of modeling both the observational causal structure, defined as the relations among the observed variables, and latent dynamic processes inherent to real-world climate systems. We understand the climate system through a causal lens and establish the identifiability guarantees for jointly recovering latent dynamics and observational causal graphs. Intuitively, the temporal structure enables leveraging contextual observable information to identify latent factors, while the inferred latent dynamics, in turn, modulate how observational causal graphs evolve. We instantiate this insight in a state-space Variational AutoEncoder (VAE), which can conduct nonparametric **Ca**usal **D**iscovery and **Re**presentation learning (**CaDRe**) simultaneously.

CaDRe employs parallel flow-based priors to *learn independent components* to reflect structural dependencies, and introduces gradient-based structural penalties on both latent transitions and decoders to ensure identifiability. Extensive synthetic experiments on the identification of latent representation learning and causal discovery validate our theoretical guarantees. On real-world climate data, CaDRe achieves competitive forecasting accuracy, indicating the effectiveness of the learned temporal process. The visualized causal graphs align with known scientific phenomena, *e.g.*, wind circulation and land–sea interactions, and further reveal structural patterns that may inspire new hypotheses in climate science.

2 Problem Setup

Technical Notations. We present the notations in a climate system, a terminology widely used in ICA literature [28]. We observed a time-series of observed variables $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$, whereas their underlying factors $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T]$ are unobservable. Regarding the system in one time-step, as depicted in Figure 1, it consists of observed variables $\mathbf{x}_t := [x_{t,i}]_{i \in \mathcal{I}}$ with index set $\mathcal{I} = \{1, 2, \dots, d_x\}$, and latent variables $\mathbf{z}_t := [z_{t,j}]_{j \in \mathcal{J}}$ indexed by $\mathcal{J} = \{1, 2, \dots, d_z\}$. Let $\mathbf{pa}(\cdot)$ denotes the parent variables, $\mathbf{pa}_O(\cdot)$ refers to observable parents, and $\mathbf{pa}_L(\cdot)$ indicates the latent parents. In particular, $\mathbf{pa}_L(\cdot)$ comprises latent variables from both the current and previous time step. Throughout the paper, the hat notation, *e.g.*, $\hat{\mathbf{x}}_t$, denotes estimated variables or functions.

Data Generating Process. Given time series data $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_T]$, we model climate evolution with the Structural Equation Model (SEM) [59]:

$$x_{t,i} = g_i(\mathbf{pa}_O(x_{t,i}), \mathbf{pa}_L(x_{t,i}), s_{t,i}), \quad z_{t,j} = f_j(\mathbf{pa}_L(z_{t,j}), \epsilon^z_{t,j}), \quad s_{t,i} = g_{s_i}(\mathbf{z}_t, \epsilon^x_{t,i}), \quad (1)$$

where g_i, f_j are differentiable, and noises $\epsilon^z_{t,j}, \epsilon^x_{t,i}$ are mutually independent. Each observed $x_{t,i}$ depends on other observed variables and latent factors $\mathbf{z}_t$ (e.g., temperature influenced by solar radiation and neighboring regions). The stochastic term $s_{t,i}$ captures variability conditioned on $\mathbf{z}_t$ (e.g., CO_2 perturbations [76]), while latent variables $\mathbf{z}_t$ evolve through both contemporaneous and time-lagged interactions. Additionally, we adopt an assumption [75] in causal discovery:

Assumption 1. *The distribution over* $(\mathbf{X}, \mathbf{Z})$ *is Markov and faithful to a Directed Acyclic Graph (DAG).*

3 Identification Theory

3.1 Latent Space Recovery and Latent Variables Identification

We consider, without loss of generality, a first-order Markov structure, in which three consecutive observations $\{\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}\}$ are used as contextual information. The generalization to higher-order Markov structures is discussed in Appendix F.1. To formalize the stochastic generation process, we first introduce an operator L [13] to represent distribution-level transformations, that is, how one probability distribution is pushed forward to another. Given two random variables a and b with supports $\mathcal{A}$ and $\mathcal{B}$ respectively, the transformation $p_a \mapsto p_b$ is formalized as:

$$p_b = L_{b|a} \circ p_a, \text{ where } L_{b|a} \circ p_a := \int_{\mathcal{A}} p_{b|a}(\cdot \mid a) p_a(a) da. \tag{2}$$

For example, operators $L_{\mathbf{x}_{t+1}|\mathbf{z}_t}$ and $L_{\mathbf{x}_{t-1}|\mathbf{x}_{t+1}}$ denote the distributional transformations $p_{\mathbf{z}_t} \mapsto p_{\mathbf{x}_{t+1}}$ and $p_{\mathbf{x}_{t+1}} \mapsto p_{\mathbf{x}_{t-1}}$. Using such operators, we address the challenge of recovering latent space under "causally-related" observations. When the observed causal graph is a DAG, information flows along causal pathways without being trapped in *self-loops*, allowing causal influence to be traced back through the reverse DAG direction via the "short reaction lag" [15]. This requires the generative operator to be *injective*, ensuring that transformations preserve full distributional information:

Lemma 1. *(Injective DAG Operator) Under Assumption 1,* $L_{\mathbf{x}_t|\mathbf{s}_t}$ *is injective for all* $t \in \mathcal{T}$.

This result shows that the nonlinear causal DAG over $\mathbf{x}_t$ does not hinder latent space recovery. A key challenge, however, is that $\mathbf{z}_t$ cannot be recovered from a single noisy $\mathbf{x}_t$, since the stochasticity makes the value-level mapping ill-posed. We therefore target identifiability at the distributional level. Crucially, neighboring observations $\mathbf{x}_{t-1}$ and $\mathbf{x}_{t+1}$ carry informative signals about $\mathbf{z}_t$ when they undergo *minimal changes*. We formalize this in the following theorem, which establishes *nonparametric* identifiability of the latent submanifold via distributional variations in context.

Theorem 1. ***(Identifiability of Latent Space)*** *Suppose observed variables and hidden variables follow the data-generating process in Eq.* (1)*, and estimated observations match the true joint distribution of* $\{\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}\}$. *The following assumptions are imposed:*

A1 (Computable Probability:) The joint, marginal, and conditional distributions of $(\mathbf{x}_t, \mathbf{z}_t)$ *are all bounded and continuous.*

A2 (Contextual Variability:) The operators $L_{\mathbf{x}_{t+1}|\mathbf{z}_t}$ *and* $L_{\mathbf{x}_{t-1}|\mathbf{x}_{t+1}}$ *are injective and bounded.*

A3 (Latent Drift:) For any $\mathbf{z}_t^{(1)}, \mathbf{z}_t^{(2)} \in \mathcal{Z}_t$ *where* $\mathbf{z}_t^{(1)} \neq \mathbf{z}_t^{(2)}$*, we have* $p(\mathbf{x}_t|\mathbf{z}_t^{(1)}) \neq p(\mathbf{x}_t|\mathbf{z}_t^{(2)})$.

A4 (Differentiability:) There exists a functional M *such that* $M\left[p_{\mathbf{x}_t|\mathbf{z}_t}(\cdot \mid \mathbf{z}_t)\right] = h_z(\mathbf{z}_t)$ *for all* $\mathbf{z}_t \in \mathcal{Z}_t$*, where* h_z *is differentiable.*

Then we have $\hat{\mathbf{z}}_t = h_z(\mathbf{z}_t)$*, where* $h_z : \mathbb{R}^{d_z} \to \mathbb{R}^{d_z}$ *is an invertible and differentiable function.*

After recovering the latent space, we aim to enhance interpretability by ensuring that each latent component corresponds to a distinct physical variable. To achieve this, we introduce a sparsity assumption on the latent dynamics, which is motivated by that physical climate factors—such as solar radiation, atmospheric pressure, or ocean currents—tend to exhibit localized sparse influences. Please refer to Appendix A.3 for the *component-wise identifiability of latent variables*.

3.2 Nonparametric Causal Discovery with the Hidden Dynamic Process

Building upon the results on recovering latent representations, we now seek to identify general nonlinear causal graphs over $\mathbf{x}_t$, even if they are modulated by a hidden dynamic process. Recent works [54, 64] extend the ICA-based Causality Discovery (CD) [72] to nonparametric settings via nonlinear ICA [31]. However, these methods are not applicable in the presence of latent confounders. To overcome this limitation, we establish a refined connection between SEMs and nonlinear ICA.

Lemma 2. *(Nonlinear SEM ⇔ Nonlinear ICA) There exists a function m_i, which is differentiable w.r.t. $s_{t,i}$ and $\mathbf{x}_t$, for any fixed $s_{t,i}$ and $\mathbf{z}_t$, such that the following two representations,*

$$x_{t,i} = g_i(\mathbf{pa}_O(x_{t,i}), \mathbf{pa}_L(x_{t,i}), s_{t,i}) \quad \textit{and} \quad x_{t,i} = m_i(\mathbf{z}_t, \mathbf{s}_t) \tag{3}$$

describe the same data-generating process. That is, both expressions yield the same value of $x_{t,i}$.

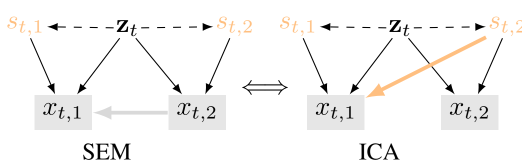


Figure 2: **Equivalent SEM and ICA.** The gray line in SEM denotes the influence $x_{t,2} \to x_{t,1}$ through the observation causal relation, which is equivalently represented as an indirect effect (the orange line): $s_{t,2} \dashrightarrow x_{t,1}$ in ICA, which can be decomposed into $s_{t,2} \to x_{t,2}$ and $x_{t,2} \to x_{t,1}$.

After establishing this equivalence, we proceed to perform CD via the nonlinear ICA with latent variables. We begin by introducing the Jacobian matrices on this data generating process, as they serve as proxies for the (nonlinear) adjacency matrix. For all $(i,j) \in \mathcal{I} \times \mathcal{I}$, we define $[\mathbf{J}_m(\mathbf{s}_t)]_{i,j} = \frac{\partial x_{t,i}}{\partial s_{t,j}}$, $[\mathbf{J}_g(\mathbf{x}_t)]_{i,j} = \frac{\partial x_{t,i}}{\partial x_{t,j}}$, and $\mathbf{D}_m(\mathbf{s}_t) = \mathrm{diag}(\frac{\partial x_{t,1}}{\partial s_{t,1}}, \frac{\partial x_{t,2}}{\partial s_{t,2}}, \ldots, \frac{\partial x_{t,d_x}}{\partial s_{t,d_x}})$, $\mathbf{I}_{d_x}$ is the identity matrix in $\mathbb{R}^{d_x \times d_x}$. Here, $\mathbf{J}_m(\mathbf{s}_t)$ corresponds to the mixing process of nonlinear ICA, as described on the R.H.S. of Eq. (3). Note that $\mathbf{J}_g(\mathbf{x}_t)$ signifies the observational causal graph in the nonlinear SEM, the L.H.S. of Eq. (3), provided the faithfulness assumption outlined below holds.

Assumption 2 (Functional Faithfulness)**.** *The causal adjacency structure among observed variables is given by the support of the Jacobian matrix $\mathbf{J}_g(\mathbf{x}_t)$.*

This assumption implies *edge minimality* in causal graphs, analogous to the structural minimality discussed in [60] (Remark 6.6) and minimality in [91], which enables us to establish a equivalence between the observational causal graph in SEM and the mixing structure in nonlinear ICA.

Theorem 2. ***(Functional Equivalence)*** *Consider the two types of data generating process described in Eq.* (3)*, the following equation always holds:*

$$\mathbf{J}_g(\mathbf{x}_t)\mathbf{J}_m(\mathbf{s}_t) = \mathbf{J}_m(\mathbf{s}_t) - \mathbf{D}_m(\mathbf{s}_t). \tag{4}$$

Building upon these SEM–ICA connections, we derive sufficient conditions under which the observational causal graph becomes identifiable in virtue of the recovered latent processes.

Theorem 3. ***(Identifiability of Observational Causal Graph)*** *Let $\mathbf{A}_{t,k} = \log p(\mathbf{s}_{t,k}|\mathbf{z}_t)$, assume that $\mathbf{A}_{t,k}$ is twice differentiable in $s_{t,k}$ and is differentiable in $z_{t,l}$, where $l = 1, 2, ..., d_z$. Suppose Assumption 1, 2 holds true, and*

A5 (Generation Variability). For any estimated $\hat{g}_m$ that makes $\mathbf{x}_t = \hat{\mathbf{x}}_t = \hat{m}(\hat{\mathbf{z}}_t, \hat{\mathbf{s}}_t)$, let

$$\mathbf{V}(t,k) := \left[\frac{\partial^2 \mathbf{A}_{t,k}}{\partial s_{t,k} \partial z_{t,1}}, \ldots, \frac{\partial^2 \mathbf{A}_{t,k}}{\partial s_{t,k} \partial z_{t,d_z}}\right], \mathbf{U}(t,k) := \left[\frac{\partial^3 \mathbf{A}_{t,k}}{\partial s_{t,k} \partial^2 z_{t,1}}, \ldots, \frac{\partial^3 \mathbf{A}_{t,k}}{\partial s_{t,k} \partial^2 z_{t,d_z}}\right]^T,$$

where for $k = 1, 2, \ldots, d_x$, $2d_x$ vector functions $\mathbf{V}(t,1), \ldots, \mathbf{V}(t,d_x), \mathbf{U}(t,1), \ldots, \mathbf{U}(t,d_x)$ are linearly independent. Then we attain ordered component-wise identifiability (Definition 4), and the structure of the observational causal graph is identifiable, i.e., supp$(\mathbf{J}_g(\mathbf{x}_t)) =$ supp$(\mathbf{J}_{\hat{g}}(\hat{\mathbf{x}}_t))$.

Based on the theoretical guarantees above, we present the **estimation methodology** in Appendix C, and provide the **experimental results** in both simulated and real-world datasets in the Appendix D.

4 Conclusion

We focused on the causal understanding of climate science and proposed a causal model with latent processes and directly causally-related observed variables. We establish identifiability results and develop an estimation approach to uncovering the latent causal variables, latent causal process, and observational causal structures from the climate system, aiming to shed light on answering "why" questions in climate. Simulated experiments validate our theoretical findings, and real-world experiments offer causal insights for climate science.

Limitations. Our method shows performance degradation as the data dimensionality increases. A potential solution is to adopt a divide-and-conquer strategy by partitioning the variables into lower-dimensional subsets using prior geographical information.

References

[1] Kashif Abbass, Muhammad Zeeshan Qasim, Huaming Song, Muntasir Murshed, Haider Mahmood, and Ijaz Younis. A review of the global climate change impacts, adaptation, and sustainable mitigation measures. *Environmental Science and Pollution Research*, 29(28):42539–42559, 2022.

[2] Lazar Atanackovic, Alexander Tong, Bo Wang, Leo J Lee, Yoshua Bengio, and Jason S Hartford. Dyngfn: Towards bayesian inference of gene regulatory networks with gflownets. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. In *International Conference on Machine Learning*, pages 899–908. PMLR, 2018.

[4] Tom Beucler and et al. Climatenet: Bringing the power of deep learning to climate science at scale. *arXiv preprint arXiv:2101.07148*, 2021.

[5] Julien Boé and Laurent Terray. Land–sea contrast, soil-atmosphere and cloud-temperature interactions: interplays and roles in future summer european climate change. *Climate dynamics*, 42(3):683–699, 2014.

[6] Philippe Brouillard, Sébastien Lachapelle, Julia Kaltenborn, Yaniv Gurwicz, Dhanya Sridhar, Alexandre Drouin, Peer Nowack, Jakob Runge, and David Rolnick. Causal representation learning in temporal data via single-parent decoding. *arXiv preprint arXiv:2410.07013*, 2024.

[7] Raymond J Carroll, Xiaohong Chen, and Yingyao Hu. Identification and estimation of nonlinear models using two samples with nonclassical measurement errors. *Journal of nonparametric statistics*, 22(4):379–399, 2010.

[8] Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, and Kun Zhang. Caring: Learning temporal causal representation under non-invertible generation process. *arXiv preprint arXiv:2401.14535*, 2024.

[9] Yi-Leng Chen and Jian-Jian Wang. The effects of precipitation on the surface temperature and airflow over the island of hawaii. *Monthly weather review*, 123(3):681–694, 1995.

[10] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[11] John B Conway. *A course in functional analysis*, volume 96. Springer Science & Business Media, 1994.

[12] Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. *arXiv preprint arXiv:2312.11001*, 2023.

[13] Nelson Dunford and Jacob T. Schwartz. *Linear Operators*. John Wiley & Sons, New York, 1971.

[14] Imme Ebert-Uphoff and Yi Deng. Causal discovery for climate research using graphical models. *Journal of Climate*, 25(17):5648–5665, 2012.

[15] Franklin M Fisher. A correspondence principle for simultaneous equation models. *Econometrica: Journal of the Econometric Society*, pages 73–92, 1970.

[16] John R Freeman. Granger causality and the times series analysis of political relationships. *American Journal of Political Science*, pages 327–358, 1983.

[17] Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems*, 33:12615–12625, 2020.

[18] Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in neural information processing systems*, 34:28233–28248, 2021.

[19] Shiyu Gu, Tim Januschowski, and Jan Gasthaus. Efficiently modeling time series with missing data using a state space approach. In *NeurIPS Time Series Workshop*, 2021.

[20] Shiyu Gu, David Salinas, Valentin Flunkert, and Jan Gasthaus. Combining latent state-space models and structural time series models for probabilistic forecasting. *International Journal of Forecasting*, 37(3):1182–1199, 2021.

[21] Shiyu Gu, David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Parameterization of state space models for forecasting with structured latent dynamics. *arXiv preprint arXiv:2202.09384*, 2022.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[23] 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.

[24] Yingyao Hu and Susanne M Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

[25] Yingyao Hu and Matthew Shum. Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics*, 171(1):32–44, 2012.

[26] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

[27] Aapo Hyvärinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10), 2023.

[28] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

[29] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

[30] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[31] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[32] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[33] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[34] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

[35] Oleksandr Klushyn and et al. Latent-space forecasting of climate variables using variational autoencoders. *arXiv preprint arXiv:2107.01227*, 2021.

[36] Lingjing Kong, Biwei Huang, Feng Xie, Eric Xing, Yuejie Chi, and Kun Zhang. Identification of nonlinear latent hierarchical models. *Advances in Neural Information Processing Systems*, 36:2010–2032, 2023.

[37] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

[38] Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024.

[39] Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

[40] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 95–104, 2018.

[41] Remi Lam and et al. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, 2022.

[42] Jan Lemeire and Dominik Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, 23:227–249, 2013.

[43] Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the identification of temporal causal representation with instantaneous dependence. In *The Thirteenth International Conference on Learning Representations*, 2025.

[44] Juan Lin. Factorizing multivariate function classes. *Advances in neural information processing systems*, 10, 1997.

[45] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. *arXiv preprint arXiv:2206.06169*, 2022.

[46] Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-tao Xia. Timebridge: Non-stationarity matters for long-term time series forecasting. *arXiv preprint arXiv:2410.04442*, 2024.

[47] 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.

[48] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

[49] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35:9881–9893, 2022.

[50] Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. *arXiv preprint arXiv:2307.02390*, 2023.

[51] Valerio Lucarini, Richard Blender, Corentin Herbert, Francesco Ragone, Salvatore Pascale, and Jeroen Wouters. Mathematical and physical ideas for climate science. *Reviews of Geophysics*, 52(4):809–859, 2014.

[52] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

[53] Lutz Mattner. Some incomplete but boundedly complete location families. *The Annals of Statistics*, pages 2158–2162, 1993.

[54] Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in artificial intelligence*, pages 186–195. PMLR, 2020.

[55] Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. *arXiv preprint arXiv:2310.15709*, 2023.

[56] Ignavier Ng, Shengyu Zhu, Zhuangyan Fang, Haoyang Li, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 424–432. SIAM, 2022.

[57] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[58] Jaideep Pathak and et al. Fourcastnet: Global medium-range weather forecasting with graph neural networks. *arXiv preprint arXiv:2202.11214*, 2022.

[59] Judea Pearl. *Causality*. Cambridge university press, 2009.

[60] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[61] Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models. *arXiv preprint arXiv:2402.09236*, 2024.

[62] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

[63] Markus Reichstein and et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

[64] Patrik Reizinger, Yash Sharma, Matthias Bethge, Bernhard Schölkopf, Ferenc Huszár, and Wieland Brendel. Jacobian-based causal discovery with nonlinear ica. *Transactions on Machine Learning Research*, 2023.

[65] Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

[66] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022.

[67] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. Pmlr, 2020.

[68] Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

[69] 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.

[70] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. In *International Journal of Forecasting*, volume 36, pages 1181–1191. Elsevier, 2020.

[71] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[72] 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.

[73] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *The Journal of Machine Learning Research*, 19(1):2639–2709, 2018.

[74] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[75] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

[76] Adolf Stips, Diego Macias, Clare Coughlan, Elisa Garcia-Gorriz, and X San Liang. On the causal structure between co2 and global temperature. *Scientific reports*, 6(1):21691, 2016.

[77] Benjamin A. Toms and Elizabeth A. Barnes. Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, 12(12), 2020.

[78] Robert Vautard, Geert Jan Van Oldenborgh, Friederike EL Otto, Pascal Yiou, Hylke De Vries, Erik Van Meijgaard, Andrew Stepek, Jean-Michel Soubeyroux, Sjoukje Philip, Sarah F Kew, et al. Human influence on european winter wind storms such as those of january 2018. *Earth System Dynamics*, 10(2):271–286, 2019.

[79] Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. Micn: Multi-scale local and global context modeling for long-term series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022.

[80] Xue Wang, Tian Zhou, Qingsong Wen, Jinyang Gao, Bolin Ding, and Rong Jin. Card: Channel aligned robust blend transformer for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2023.

[81] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

[82] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

[83] Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with \$10k\$ parameters. In *The Twelfth International Conference on Learning Representations*, 2024.

[84] Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying causal representation learning with dynamical systems for science. *arXiv preprint arXiv:2405.13888*, 2024.

[85] Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.

[86] Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022.

[87] Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

[88] Weiwei Ye, Songgaojun Deng, Qiaosha Zou, and Ning Gui. Frequency adaptive normalization for non-stationary time series forecasting. *arXiv preprint arXiv:2409.20371*, 2024.

[89] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

[90] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

[91] Jiji Zhang. A comparison of three occam's razors for markovian causal models. *The British journal for the philosophy of science*, 2013.

[92] Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

[93] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

[94] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

[95] 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.

[96] Yujia Zheng, Ignavier Ng, Yewen Fan, and Kun Zhang. Generalized precision matrix for scalable estimation of nonparametric markov networks. *arXiv preprint arXiv:2305.11379*, 2023.

[97] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022.

[98] Yujia Zheng and Kun Zhang. Generalizing nonlinear ica beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36:13326–13355, 2023.

[99] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

Supplement to

"Learning General Causal Structures with Hidden Dynamic Process for Climate Analysis"

Appendix organization:

A Theorem Proofs

A.1 Notation List

This section collects the notations used in the theorem proofs for clarity and consistency.

Table A1: List of notations, explanations, and corresponding values.

Index	Explanation	Support
d_x	number of observed variables	$d_x \in \mathbb{N}^+$
d_z	number of latent variables	$d_z \in \mathbb{N}^+$ and $d_z \leq d_x$
t	time index	$t \in \mathbb{N}^+$ and $t \geq 3$
$\mathcal{I}$	index set of observed variables	$\mathcal{I} = \{1, 2, \ldots, d_x\}$
$\mathcal{J}$	index set of latent variables	$\mathcal{J} = \{1, 2, \ldots, d_z\}$
Variable		
$\mathcal{X}_t$	support of observed variables in time-index t	$\mathcal{X}_t \subseteq \mathbb{R}^{d_x}$
$\mathcal{Z}_t$	support of latent variables	$\mathcal{Z}_t \subseteq \mathbb{R}^{d_z}$
$\mathbf{x}_t$	observed variables in time-index t	$\mathbf{x}_t \in \mathcal{X}_t$
$\mathbf{z}_t$	latent variables in time-index t	$\mathbf{z}_t \in \mathcal{Z}_t$
$\mathbf{s}_t$	dependent noise of observations in time-index t	$\mathbf{s}_t \in \mathbb{R}^{d_x}$
$\boldsymbol{\epsilon}_{\mathbf{x}_t}$	independent noise for generating $\mathbf{s}_t$ in time-index t	$\boldsymbol{\epsilon}_{\mathbf{x}_t} \sim p_{\epsilon_x}$
$\boldsymbol{\epsilon}_{\mathbf{z}_t}$	independent noise of latent variables in time-index t	$\boldsymbol{\epsilon}_{\mathbf{z}_t} \sim p_{\epsilon_z}$
$\mathbf{z}_{t\setminus[i,j]}$	latent variables except for $z_{t,i}$ and $z_{t,j}$ in time-index t	/
Function		
$p_{a\mid b}(\cdot \mid b)$	density function of a given b	/
$p_{a,b\mid c}(a, \cdot \mid c)$	joint density function of (a, b) given a and c	/
$\mathbf{pa}(\cdot)$	variable's parents	/
$\mathbf{pa}_O(\cdot)$	variable's parents in observed space	/
$\mathbf{pa}_L(\cdot)$	variable's parents in latent space	/
$g(\cdot)$	generating function of SEM from $(\mathbf{z}_t, \mathbf{s}_t, \mathbf{x}_t)$ to $\mathbf{x}_t$	$\mathbb{R}^{d_z+2d_x} \to \mathbb{R}^{d_x}$
$m(\cdot)$	mixing function of ICA from $(\mathbf{z}_t, \mathbf{s}_t)$ to $\mathbf{x}_t$	$\mathbb{R}^{d_z+d_x} \to \mathbb{R}^{d_x}$
$h_z(\cdot)$	invertible transformation from $\mathbf{z}_t$ to $\hat{\mathbf{z}}_t$	$\mathbb{R}^{d_z} \to \mathbb{R}^{d_z}$
$\pi(\cdot)$	permutation function	$\mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$
$\text{supp}(\cdot)$	support matrix of Jacobian matrix	$\mathbb{R}^{d_x \times d_x} \to \{0, 1\}^{d_x \times d_x}$
Symbol		
$A \to B$	A causes B directly	/
$A \dashrightarrow B$	A causes B indirectly	/
$\mathbf{J}_g(\mathbf{x}_t)$	Jacobian matrix representing observed causal DAG	$\mathbf{J}_g(\mathbf{x}_t) \in \mathbb{R}^{d_x \times d_x}$
$\mathbf{J}_g(\mathbf{x}_t, \mathbf{s}_t)$	Jacobian matrix representing mixing structure from $(\mathbf{x}_t, \mathbf{s}_t)$ to $\mathbf{x}_t$	$\mathbf{J}_g(\mathbf{x}_t, \mathbf{s}_t) \in \mathbb{R}^{d_x \times d_x}$
$\mathbf{J}_m(\mathbf{s}_t)$	Jacobian matrix representing mixing structure from $\mathbf{s}_t$ to $\mathbf{x}_t$	$\mathbf{J}_m(\mathbf{s}_t) \in \mathbb{R}^{d_x \times d_x}$
$\mathbf{J}_r(\mathbf{z}_{t-1})$	Jacobian matrix representing latent time-lagged structure	$\mathbf{J}_r(\mathbf{z}_{t-1}) \in \mathbb{R}^{d_z \times d_z}$
$\mathbf{J}_r(\mathbf{z}_t)$	Jacobian matrix representing instantaneous latent causal graph	$\mathbf{J}_r(\mathbf{z}_t) \in \mathbb{R}^{d_z \times d_z}$

A.2 Proof of Theorem 1

We first introduce another operator to represent the point-wise distributional multiplication. To maintain generality, we denote two variables as a and b, with respective support sets $\mathcal{A}$ and $\mathcal{B}$.

Definition 1. *(Diagonal Operator) Consider two random variable a and b, density functions p_a and p_b are defined on some support $\mathcal{A}$ and $\mathcal{B}$, respectively. The diagonal operator $D_{b|a}$ maps the density function p_a to another density function $D_{b|a} \circ p_a$ defined by the pointwise multiplication of the function $p_{b|a}$ at a fixed point b:*

$$p_{b|a}(b \mid \cdot)p_a = D_{b|a} \circ p_a, \textit{where } D_{b|a} = p_{b|a}(b \mid \cdot). \tag{A1}$$

Proof. $\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}$ are conditional independent given $\mathbf{z}_t$, which implies two equations:

$$p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{z}_t) = p(\mathbf{x}_{t-1} \mid \mathbf{z}_t), \quad p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t) = p(\mathbf{x}_{t+1} \mid \mathbf{z}_t). \tag{A2}$$

We can obtain $p(\mathbf{x}_{t+1}, \mathbf{x}_t \mid \mathbf{x}_{t-1})$ directly from the observations, $p(\mathbf{x}_{t-1})$ and $p(\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{x}_{t-1})$, and then the transformation in density function are established by

$$
\begin{aligned}
p(\mathbf{x}_{t+1}, \mathbf{x}_t \mid \mathbf{x}_{t-1}) &= \underbrace{\int_{\mathcal{Z}_t} p(\mathbf{x}_{t+1}, \mathbf{x}_t, \mathbf{z}_t \mid \mathbf{x}_{t-1}) d\mathbf{z}_t}_{\text{integration over } \mathcal{Z}_t} = \underbrace{\int_{\mathcal{Z}_t} p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{z}_t, \mathbf{x}_{t-1}) p(\mathbf{x}_t, \mathbf{z}_t \mid \mathbf{x}_{t-1}) d\mathbf{z}_t}_{\text{factorization of joint conditional probability}} \\
&= \underbrace{\int_{\mathcal{Z}_t} p(\mathbf{x}_{t+1} \mid \mathbf{z}_t) p(\mathbf{x}_t, \mathbf{z}_t \mid \mathbf{x}_{t-1}) d\mathbf{z}_t}_{\text{by } p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t) = p(\mathbf{x}_{t+1} \mid \mathbf{z}_t)} = \underbrace{\int_{\mathcal{Z}_t} p(\mathbf{x}_{t+1} \mid \mathbf{z}_t) p(\mathbf{x}_t \mid \mathbf{z}_t) p(\mathbf{z}_t \mid \mathbf{x}_{t-1}) d\mathbf{z}_t}_{\text{by } p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{z}_t) = p(\mathbf{x}_{t-1} \mid \mathbf{z}_t)} .
\end{aligned}
\tag{A3}
$$

Then we show how to transform the Eq. (A3) to the form of spectral decomposition:

$$
\begin{aligned}
\Longrightarrow \quad & \int_{\mathcal{X}_{t-1}} p(\mathbf{x}_{t+1}, \mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}) d\mathbf{x}_{t-1} = \\
& \qquad \int_{\mathcal{X}_{t-1}} \int_{\mathcal{Z}_t} p(\mathbf{x}_{t+1} \mid \mathbf{z}_t) p(\mathbf{x}_t \mid \mathbf{z}_t) p(\mathbf{z}_t \mid \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}) d\mathbf{z}_t d\mathbf{x}_{t-1} && \text{(A4)} \\
\Longrightarrow \quad & [L_{\mathbf{x}_t; \mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}} p](\mathbf{x}_{t+1}) = [L_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} D_{\mathbf{x}_t \mid \mathbf{z}_t} L_{\mathbf{z}_t \mid \mathbf{x}_{t-1}} p](\mathbf{x}_{t+1}), && \text{(A5)} \\
\Longrightarrow \quad & L_{\mathbf{x}_t; \mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}} = L_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} D_{\mathbf{x}_t \mid \mathbf{z}_t} L_{\mathbf{z}_t \mid \mathbf{x}_{t-1}} && \text{(A6)} \\
\Longrightarrow \quad & \int_{\mathbf{x}_t \in \mathcal{X}_t} L_{\mathbf{x}_t; \mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}} \, d\mathbf{x}_t = \int_{\mathbf{x}_t \in \mathcal{X}_t} L_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} D_{\mathbf{x}_t \mid \mathbf{z}_t} L_{\mathbf{z}_t \mid \mathbf{x}_{t-1}} \, d\mathbf{x}_t && \text{(A7)} \\
\Longrightarrow \quad & L_{\mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}} = L_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} L_{\mathbf{z}_t \mid \mathbf{x}_{t-1}} && \text{(A8)} \\
\Longrightarrow \quad & L^{-1}_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} L_{\mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}} = L_{\mathbf{z}_t \mid \mathbf{x}_{t-1}} && \text{(A9)} \\
\Longrightarrow \quad & L_{\mathbf{x}_t; \mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}} = L_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} D_{\mathbf{x}_t \mid \mathbf{z}_t} L^{-1}_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} L_{\mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}} && \text{(A10)} \\
\Longrightarrow \quad & L_{\mathbf{x}_t; \mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}} L^{-1}_{\mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}} = L_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} D_{\mathbf{x}_t \mid \mathbf{z}_t} L^{-1}_{\mathbf{x}_{t+1} \mid \mathbf{z}_t}. && \text{(A11)} \\
\Longrightarrow \quad & L_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} D_{\mathbf{x}_t \mid \mathbf{z}_t} L^{-1}_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} = (C L_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} P)(P^{-1} D_{\mathbf{x}_t \mid \mathbf{z}_t} P)(P^{-1} L^{-1}_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} C^{-1}) && \text{(A12)} \\
\Longrightarrow \quad & L_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} = C L_{\mathbf{x}_{t+1} \mid \hat{\mathbf{z}}_t} P, \quad D_{\mathbf{x}_t \mid \mathbf{z}_t} = P^{-1} D_{\mathbf{x}_t \mid \hat{\mathbf{z}}_t} P && \text{(A13)}
\end{aligned}
$$

where

- in Eq. (A4), we add the integration over $\mathcal{X}_{t-1}$ in both sides of Eq. (A3). s
- in Eq. (A5), we replace the probability with operators by using Eq. (2) and Definition 1. Specifically, we have: $L_{\mathbf{x}_t; \mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}} = \int_{\mathcal{X}_{t-1}} p_{\mathbf{x}_{t+1}}(\mathbf{x}_t, \cdot \mid \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}) d\mathbf{x}_{t-1}$.
- in Eq. (A9), the operator $L_{\mathbf{x}_{t+1} \mid \mathbf{z}_t}$ is injective by Assumption 1
- in Eq. (A10), the $L_{\mathbf{z}_t \mid \mathbf{x}_{t-1}}$ in Eq. (A6) is substituted by Eq. (A9):
- in Eq. (A11), if $L_{\mathbf{x}_{t-1} \mid \mathbf{x}_{t+1}}$ is injective, then $L^{-1}_{\mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}}$ exists and is densely defined over $\mathcal{F}(\mathcal{X}_{t+1})$.
- in Eq. (A13), Assumption 1 ensures that $L_{\mathbf{x}_t; \mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}} L^{-1}_{\mathbf{x}_{t+1} \mid \mathbf{x}_{t-1}}$ is bounded; by the uniqueness of spectral decomposition (see e.g., [11] Ch. VII and [13] Theorem XV 4.5), $L_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} D_{\mathbf{x}_t \mid \mathbf{z}_t} L^{-1}_{\mathbf{x}_{t+1} \mid \mathbf{z}_t}$ admits a unique spectral decomposition in which the eigenvalues, *i.e.*, $D_{\mathbf{x}_t \mid \mathbf{z}_t}$, which are precisely the entries of $\{p_{\mathbf{x}_t \mid \mathbf{z}_t}(\mathbf{x}_t \mid \mathbf{z}_t)\}$, and eigenfunctions, *i.e.*, $D_{\mathbf{x}_t \mid \mathbf{z}_t}$, which columns are $\{p_{\mathbf{x}_{t+1} \mid \mathbf{z}_t}(\cdot \mid \mathbf{z}_t)\}$, up to standard indeterminacies. C is an nonzero scalar rescaling eigenvalues, and P is a operator permuting the eigenvalues and eigenfunctions.

We obtain a unique spectral decomposition in Eq. (A13) with permutation and scaling indeterminacies. In the following, we will show how these indeterminacies can be resolved—if not, what informative results can still be inferred.

First, considering the arbitrary scaling C, since the normalizing condition

$$
\int_{\mathcal{X}_{t+1}} p_{\mathbf{x}_{t+1} \mid \hat{\mathbf{z}}_t} \, d\mathbf{x}_{t+1} = 1 \tag{A14}
$$

must hold for every $\hat{\mathbf{z}}_t$, one only solution of $\int_{\mathcal{X}_{t+1}} C p_{\mathbf{x}_{t+1} \mid \mathbf{z}_t} \, d\mathbf{x}_{t+1} = 1$ is to set $C = 1$.

Second, regarding the permutation indeterminacy, we start from $D_{\mathbf{x}_t|\mathbf{z}_t} = P^{-1} D_{\mathbf{x}_t|\hat{\mathbf{z}}_t} P$. The operator, $D_{\mathbf{x}_t|\mathbf{z}_t}$, corresponding to the set $\{p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid \mathbf{z}_t)\}$ for fixed $\mathbf{x}_t$ and all $\mathbf{z}_t$, admits a unique solution (P only change the entry position):

$$\{p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid \mathbf{z}_t)\} = \{p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\mathbf{x}_t \mid \hat{\mathbf{z}}_t)\}, \quad \text{for all } \mathbf{z}_t, \hat{\mathbf{z}}_t \tag{A15}$$

Due to the set is unorder, the only way to match the R.H.S. with the L.H.S. in a consistent order is to exchange the conditioning variables, that is,

$$\{p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid \mathbf{z}_t^{(1)}), p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid \mathbf{z}_t^{(2)}), \ldots\} = \{p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\mathbf{x}_t \mid \hat{\mathbf{z}}_t^{(1)}), p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\mathbf{x}_t \mid \hat{\mathbf{z}}_t^{(2)}), \ldots\} \tag{A16}$$

$$\implies \quad [p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid \mathbf{z}_t^{(\pi(1))}), p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid \mathbf{z}_t^{(\pi(2))}), \ldots] = [p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\mathbf{x}_t \mid \hat{\mathbf{z}}_t^{(\pi(1))}), p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\mathbf{x}_t \mid \hat{\mathbf{z}}_t^{(\pi(2))}), \ldots] \tag{A17}$$

where superscript $(\cdot)$ denotes the index of a conditioning variable, and π is reindexing the conditioning variables. We use a relabeling map h to represent its corresponding value mapping:

$$p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid h(\mathbf{z}_t)) = p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\mathbf{x}_t \mid \hat{\mathbf{z}}_t), \quad \text{for all } \mathbf{z}_t, \hat{\mathbf{z}}_t. \tag{A18}$$

By Assumption 1, different $\mathbf{z}_t$ corresponds to different $p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid \mathbf{z}_t)$, there is no repeated element in $\{p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid \mathbf{z}_t)\}$ (and $\{p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\mathbf{x}_t \mid \hat{\mathbf{z}}_t)\}$). Hence, the relabelling map h is one-to-one (invertible). Furthermore, Assumption 4 implies that $p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid h(\mathbf{z}_t))$ determines a unique $h(\mathbf{z}_t)$. The same holds for the $p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\mathbf{x}_t \mid \hat{\mathbf{z}}_t)$, implying that

$$p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid h(\mathbf{z}_t)) = p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\mathbf{x}_t \mid \hat{\mathbf{z}}_t) \implies \hat{\mathbf{z}}_t = h(\mathbf{z}_t). \tag{A19}$$

Next, Assumption 1 implies that the function h must be differentiable. Since the VAE is differentiable, we can learn a differentiable function h that satisfies Assumption 1. Consider $\hat{\mathbf{z}}_t$ related to $\mathbf{z}_t$ via $\hat{\mathbf{z}}_t = h(\mathbf{z}_t)$. Then, we have

$$M\left[p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\cdot \mid \mathbf{z}_t)\right] = M\left[p_{\mathbf{x}_t|\mathbf{z}_t}(\cdot \mid h(\mathbf{z}_t))\right] = h(\mathbf{z}_t), \tag{A20}$$

which is equal to $\hat{\mathbf{z}}_t$ only if h is differentiable.
To ensure the latent dimension d_z is also identifiable, we analyze two scenarios :

i. $d_{\hat{z}} > d_z$: d_z latent components in $\hat{\mathbf{z}}_t$ are sufficient to explain $\mathbf{x}_t$, *i.e.*,

$$p(\mathbf{x}_t \mid \mathbf{z}_{t,:d_{\hat{z}}-d_z}, \mathbf{z}^{(1)}_{t,d_{\hat{z}}-d_z:}) = p(\mathbf{x}_t \mid \mathbf{z}_{t,:d_{\hat{z}}-d_z}, \mathbf{z}^{(2)}_{t,d_{\hat{z}}-d_z:}), \tag{A21}$$

which contradicts the Assumption 1.

ii. $d_{\hat{z}} < d_z$: This suggests that only $d_{\hat{z}}$ dimensions are sufficient to reconstruct $\mathbf{x}_t$, leaving $d_z - d_{\hat{z}}$ components constant, which violates that there are d_z latent *variables*.

□

More Discussions of Assumption 1 The injectivity of the operator enables us to take inverses of certain operators, which is commonly made in nonparametric identification [24, 7, 25]. Intuitively, different input distribution correpsonds to different output distribution. In the context of the climate system, it represent the necessity of temporal variability. However, it is difficult to formalize it in terms of functions. We give some examples in terms of $p_a \Rightarrow p_b$ to make it understandable:

Example 1. $b = g(a)$, where g is an invertible function.

Example 2. $b = a + \epsilon$, where $p(\epsilon)$ must not vanish everywhere after the Fourier transform (Theorem 2.1 in [53]).

Example 3. $b = g(a) + \epsilon$, where the same conditions from Examples 1 and 2 are required.

Example 4. $b = g_1(g_2(a) + \epsilon)$, a post-nonlinear model with invertible nonlinear functions g_1, g_2, combining the assumptions in **Examples 1-3**.

Example 5. $b = g(a, \epsilon)$, where the joint distribution $p(a, b)$ follows an exponential family.

Example 6. $b = g(a, \epsilon)$, a general nonlinear formulation. Certain deviations from the nonlinear additive model (**Example 3**), e.g., polynomial perturbations, can still be tractable.

A.3 Component-Wise Identifiability of Latent Variables

Theorem A1. *(Component-Wise Identifiability of Latent Variables [43]) Let $\mathbf{c}_t \triangleq \{\mathbf{z}_{t-1}, \mathbf{z}_t\}$ and $\mathcal{M}_{\mathbf{c}_t}$ be the variable set of two consecutive timestamps and the corresponding Markov network, respectively. Suppose the following assumptions hold:*

i. *(Smooth and Positive Density): The probability function of the latent variables $\mathbf{c}_t$ is smooth and positive,* i.e.*, $p_{\mathbf{c}_t}$ is third-order differentiable and $p_{\mathbf{c}_t} > 0$ over $\mathbb{R}^{2n}$.*

ii. *(Sufficient Variability)s: Denote $|\mathcal{M}_{\mathbf{c}_t}|$ as the number of edges in Markov network $\mathcal{M}_{\mathbf{c}_t}$. Let*

$$
\begin{aligned}
w(m) = &\left(\frac{\partial^3 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,1}^2 \partial z_{t-2,m}}, \ldots, \frac{\partial^3 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,2n}^2 \partial z_{t-2,m}} \right) \oplus \\
&\left(\frac{\partial^2 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,1} \partial z_{t-2,m}}, \ldots, \frac{\partial^2 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,2n} \partial z_{t-2,m}} \right) \oplus \left(\frac{\partial^3 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,i} \partial c_{t,j} \partial z_{t-2,m}} \right)_{(i,j) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})},
\end{aligned}
\tag{A22}
$$

where $\oplus$ denotes the concatenation operation and $(i,j) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})$ denotes all pairwise indices such that $c_{t,i}, c_{t,j}$ are adjacent in $\mathcal{M}_{\mathbf{c}_t}$. For $m \in \{1, \ldots, n\}$, there exist $4n + 2|\mathcal{M}_{\mathbf{c}_t}|$ different values of $\mathbf{z}_{t-2,m}$ as the $4n + 2|\mathcal{M}_{\mathbf{c}_t}|$ values of vector functions $w(m)$ are linearly independent.

iii. *(Sparse Latent Process): For any $z_{t,i} \in \mathbf{z}_t$, the intimate neighbor set of $z_{t,i}$ is an empty set, where the intimate neighbor set is defined as*

Definition 2. *(Intimate Neighbor Set) Consider a Markov network $\mathcal{M}_Z$ over variables set Z, and the intimate neighbor set of variable $z_{t,i}$ is*

$$
\Psi_{\mathcal{M}_{\mathbf{c}_t}}(c_{t,i}) \triangleq \left\{ c_{t,j} \;\middle|\; \begin{array}{c} c_{t,j} \text{ is adjacent to } c_{t,i} \text{ and also adjacent} \\ \text{to all other neighbors of } c_{t,i},\ c_{t,j} \in \mathbf{c}_t \setminus \{c_{t,i}\} \end{array} \right\}
\tag{A23}
$$

iv. *(Transition Variability): For any pair of adjacent latent variables $z_{t,i}, z_{t,j}$ at time step t, their time-delayed parents are not identical,* i.e.*, $\mathbf{pa}(z_{t,i}) \neq \mathbf{pa}(z_{t,j})$.*

*Then for any two different entries $\hat{c}_{t,k}, \hat{c}_{t,l} \in \hat{\mathbf{c}}_t$ that are **not adjacent** in the Markov network $\mathcal{M}_{\hat{\mathbf{c}}_t}$ over estimated $\hat{\mathbf{c}}_t$,*

i. *The estimated Markov network $\mathcal{M}_{\hat{\mathbf{c}}_t}$ is isomorphic to the ground-truth Markov network $\mathcal{M}_{\mathbf{c}_t}$.*

ii. *There exists a permutation π of the estimated latent variables, such that $z_{t,i}$ and $\hat{z}_{t,\pi(i)}$ is one-to-one corresponding,* i.e.*, $z_{t,i}$ is component-wise identifiable.*

iii. *The causal graph of the latent causal process is identifiable.*

Proof Sketch and Discussions. Once latent space is recovered by Theorem 1, *i.e.*, *(i)* $\hat{\mathbf{z}}_t = h_z(\mathbf{z}_t)$ is established, leveraging two properties of latent space—namely, *(ii)* the sparsity in the latent Markov network, and *(iii)* $z_{t,i} \perp\!\!\!\perp z_{t,j} \mid \mathbf{z}_{t-1}, \mathbf{z}_{t/[i,j]}$ if $z_{t,i}, z_{t,j}$ $(i \neq j)$ are not adjacent in Markov network, we can obtain the component-wise identifiability of latent variables under *sufficient variability* assumption. In contrast to prior work on CRL with component-wise identifiability [94, 43], which typically requires the full invertible mapping g, and assume multiple distributions or temporal steps while only uses a single measurement for the latent space recovery, our approach fully exploits the temporally adjacent measurements, thereby avoiding the need for strong assumption.

Furthermore, we show that using three adjacent time steps, including future observations rather than relying solely on the past, suffices to recover the entire latent process. This temporal window matches that in Theorem 1. Having established the required conditions, we can directly apply the identifiability result from [43] to complete the proof of the identifiable latent causal process.

A.4 Proof of Lemma 2

Definition 3. *(Causal Order) $x_{t,i}$ is in the τ-th causal order if only observed variables in the $(\tau-1)$-th causal order directly influence it. We specify $\mathbf{z}_t$ is in the 0-th causal order.*

For an observed variable $x_{t,i}$, we define the set $\mathcal{P}$ to include all variables in $\mathbf{x}_t$ involved in generating $x_{t,i}$, initialized as $\mathcal{P} = \mathbf{pa}_O(x_{t,i})$. The upper bound of the cardinality of $\mathcal{P}$ is given by $\mathcal{U}(|\mathcal{P}|)$, which satisfies $\mathcal{U}(|\mathcal{P}|) = d_x - 1$ initially. Let $\mathcal{Q}$ denote the set of latent variables, and define the separated set as $\mathcal{S}$, where $g_{s_i}(\mathbf{pa}_L(x_{t,i}), \epsilon_{x_{t,i}})$ is denoted by $s_{t,i}$. Initially, $\mathcal{S} = \{s_{t,i}\}$. We express $x_{t,i}$ as $x_{t,i} = g_i(\mathcal{P}, \mathcal{S}, \mathcal{Q})$, and traverse all $x_{t,j} \in \mathbf{x}_t$ in descending causal order τ_j, performing the following operations:

i. Remove $x_{t,j}$ from $\mathcal{P}$ and apply Eq. (1) to obtain

$$x_{t,i} = f_1\left(\mathcal{P} \setminus \{x_{t,j}\}, \mathcal{S}, \mathcal{Q}, \mathbf{pa}_O(x_{t,j}), \mathbf{pa}_L(x_{t,j}), s_{t,j}\right). \tag{A24}$$

Then, update $\mathcal{P} \leftarrow (\mathcal{P} \setminus \{x_{t,j}\}) \cup \mathbf{pa}_O(x_{t,j})$ and $\mathcal{Q} \leftarrow \mathcal{Q} \cup \mathbf{pa}_L(x_{t,j})$. By Assumption 1, $x_{t,j}$ cannot reappear in the set of its ancestors, resulting in $\mathcal{U}(|\mathcal{P}|) \leftarrow \mathcal{U}(|\mathcal{P}|) - 1$.

ii. Assumption 1 also ensures that a variable with a lower causal order does not appear in the generation of its descendants. Hence, $x_{t,j}$ cannot appear in the generation of its descendants, since their causal orders are larger than τ_j. Similarly, $s_{t,j}$, which is involved in generating $x_{t,j}$, does not appear in the generation of its descendants. Thus, $s_{t,j} \notin \mathcal{S}$. Define the new separated set as $\mathcal{S} \leftarrow \mathcal{S} \cup \{s_{t,j}\}$, giving

$$x_{t,i} = f_2\left(\mathcal{P}, \mathcal{S}, \mathcal{Q}\right), \tag{A25}$$

where the new cardinality is updated as $|\mathcal{S}| \leftarrow |\mathcal{S}| + 1$.

Given that $\mathcal{U}(|\mathcal{P}|) \geq |\mathcal{P}|, \mathcal{U}(|\mathcal{P}|)$ ensures that this iterative process can be performed until $|\mathcal{P}| = 0$. According to the definition of data generating process, all the aforementioned functions are partially differentiable w.r.t. $\mathbf{s}_t$ and $\mathbf{x}_t$, or they are compositions of such functions. As a result, $\mathcal{Q} = \mathbf{an}_{\mathbf{z}_t}(x_{t,i})$, and there exists a function g_{m_i} such that

$$x_{t,i} = g_{m_i}(\mathbf{an}_{\mathbf{z}_t}(x_{t,i}), \mathbf{s}_t).$$

Moreover, we observe that $\mathbf{s}_t$ is in fact the ancestors $\mathbf{an}_{\epsilon_{\mathbf{x}_t}}(x_{t,i}) = \{\epsilon_{x_{t,j}} \mid s_{t,j} \in \mathcal{S}\}$, which are implied in this derivation process since $\epsilon_{x_{t,j}}$ is in one-to-one correspondence with $s_{t,j}$ through indexing.

A.5 Proof of Theorem 2

Considering the mixing function m, and the functional relation $s_{t,j} \rightarrow x_{t,i}$, corresponding $[\mathbf{J}_m(\mathbf{s}_t)]_{i,j}$, where i, j indicates the row and column index of the Jacobian matrix, respectively.

For the elements $i \neq j$**:** If there is a directed functional relationship $x_{t,j} \rightarrow x_{t,i}$, the corresponding element of the Jacobian matrix is $\frac{\partial x_{t,i}}{\partial x_{t,j}}$. If the relationship is indirect: $x_{t,j} \dashrightarrow x_{t,i}$, then for each $x_{t,k} \in \mathbf{pa}_O(x_{t,i})$, there must exist either an indirect-direct path $x_{t,j} \dashrightarrow x_{t,k} \rightarrow x_{t,i}$ or a direct-direct path $x_{t,j} \rightarrow x_{t,k} \rightarrow x_{t,i}$. In summary, through the chain rule, we obtain

$$[\mathbf{J}_m(\mathbf{s}_t)]_{i,j} = \sum_{x_{t,k} \in \mathbf{pa}_O(x_{t,i})} \frac{\partial x_{t,i}}{\partial x_{t,k}} \cdot \frac{\partial x_{t,k}}{\partial s_{t,j}}. \tag{A26}$$

For each $x_{t,k} \notin \mathbf{pa}_O(x_{t,i})$, $\frac{\partial x_{t,i}}{\partial x_{t,k}} = 0$, Eq. (A26) could be rewritten as

$$\begin{aligned}
[\mathbf{J}_m(\mathbf{s}_t)]_{i,j} &= \sum_{x_{t,k} \in \mathbf{pa}_O(x_{t,i})} \frac{\partial x_{t,i}}{\partial x_{t,k}} \cdot \frac{\partial x_{t,k}}{\partial s_{t,j}} + \sum_{x_{t,k} \notin \mathbf{pa}_O(x_{t,i})} \frac{\partial x_{t,i}}{\partial x_{t,k}} \cdot \frac{\partial x_{t,k}}{\partial s_{t,j}} \\
&= \sum_{k=1}^{d_x} \frac{\partial x_{t,i}}{\partial x_{t,k}} \cdot \frac{\partial x_{t,k}}{\partial s_{t,j}} = \sum_{k=1}^{d_x} [\mathbf{J}_g(\mathbf{x}_t)]_{i,k} \cdot [\mathbf{J}_m(\mathbf{s}_t)]_{k,j}.
\end{aligned} \tag{A27}$$

For the elements $i = j$**:** For each $x_{t,k} \in \mathbf{pa}_O(x_{t,i})$, DAG structure ensures that $x_{t,i}$ does not appear in the set of ancestors of itself. Consequently, due to the one-to-one correspondence between $s_{t,k}$ and $x_{t,i}$, we also have that $\frac{\partial x_{t,k}}{\partial s_{t,i}} = 0$. Thus, we obtain

$$[\mathbf{J}_m(\mathbf{s}_t)]_{i,i} = \frac{\partial x_{t,i}}{\partial s_{t,i}} + 0 = \frac{\partial x_{t,i}}{\partial s_{t,i}} + \sum_{k=1}^{d_x} [\mathbf{J}_g(\mathbf{x}_t)]_{i,k} \cdot [\mathbf{J}_m(\mathbf{s}_t)]_{k,i}. \tag{A28}$$

Since for $k = i$, it holds that $[\mathbf{J}_g(\mathbf{x}_t)]_{i,k} = 0$, and for $k \neq i$, we have $[\mathbf{J}_m(\mathbf{s}_t)]_{k,i} = 0$. Defining $\mathbf{D}_m(\mathbf{s}_t) = \mathrm{diag}(\frac{\partial x_{t,1}}{\partial s_{t,1}}, \ldots, \frac{\partial x_{t,d_x}}{\partial s_{t,d_x}})$, we can summarize the result as

$$\mathbf{J}_g(\mathbf{x}_t)\mathbf{J}_m(\mathbf{s}_t) = \mathbf{J}_m(\mathbf{s}_t) - \mathbf{D}_m(\mathbf{s}_t). \tag{A29}$$

A.6 Proof of Corollary A1

We establish two results that strengthen the SEM–ICA connection by relaxing modeling assumptions and enabling its practical application within generative models.

Corollary A1. *Under Assumption 1, given any $\mathbf{z}_t \in \mathcal{Z}_t$, $\mathbf{J}_m(\mathbf{s}_t)$ is a invertible matrix.*

This result unveils that the DAG structure among observed variables implies the invertibility of the mixing function m in the nonlinear ICA. As a direct consequence, by left-multiplying both sides of Eq. (4) with $\mathbf{J}_m^{-1}(\mathbf{s}_t)$, we obtain the following expression:

Corollary A2. *Observational causal graphs are represented by $\mathbf{J}_g(\mathbf{x}_t) = \mathbf{I}_{d_x} - \mathbf{D}_m(\mathbf{s}_t)\mathbf{J}_m^{-1}(\mathbf{s}_t)$.*

Proof. Eq. (A29) states that

$$(\mathbf{I}_{d_x} - \mathbf{J}_g(\mathbf{x}_t))\mathbf{J}_m(\mathbf{s}_t) = \mathbf{D}_m(\mathbf{s}_t). \tag{A30}$$

From the DAG structure specified in Condition 1 and the functional faithfulness assumption in Assumption 2, the Jacobian matrix $\mathbf{J}_g(\mathbf{x}_t)$ can be permuted into a lower triangular form via identical row and column permutations. Thus, the matrix $\mathbf{I}_{d_x} - \mathbf{J}_g(\mathbf{x}_t)$ is invertible for all $\mathbf{x}_t \in \mathcal{X}_t$. Since $\mathbf{D}_m(\mathbf{s}_t)$ is obtained via multiplication with $(\mathbf{I}_{d_x} - \mathbf{J}_g(\mathbf{x}_t))$, it follows that

$$(\mathbf{I}_{d_x} - \mathbf{J}_g(\mathbf{x}_t))^{-1}\mathbf{D}_m(\mathbf{s}_t) \tag{A31}$$

is well-defined and invertible. This, in turn, implies that $\mathbf{J}_m(\mathbf{s}_t)$ is invertible. Furthermore, we establish that

$$\text{supp}\,(\mathbf{I}_{d_x} - \mathbf{J}_g(\mathbf{x}_t)) = \text{supp}\,(\mathbf{J}_g(\mathbf{x}_t, \mathbf{s}_t)) \tag{A32}$$

since the diagonal entries of $\mathbf{J}_g(\mathbf{x}_t, \mathbf{s}_t)$ are nonzero. Given that $\mathbf{J}_g(\mathbf{x}_t, \mathbf{s}_t)$ inherits the lower triangular structure after permutation, it must also be invertible.

A.7 Proof of Theorem 3

We present some useful definitions and lemmas in our proof.

Definition 4. *(Ordered Component-wise Identifiability) Variables $\mathbf{s}_t \in \mathbb{R}^{d_x}$ and $\hat{\mathbf{s}}_t \in \mathbb{R}^{d_x}$ are identified component-wise if $\hat{s}_{t,i} = h_{s_i}(s_{t,\pi(i)})$ with invertible function h_{s_i} and $\pi(i) = i$.*

Lemma 3 (Lemma 1 in LiNGAM [72])**.** *Assume $\mathbf{M}$ is lower triangular and all diagonal elements are non-zero. A permutation of rows and columns of $\mathbf{M}$ has only non-zero entries in the diagonal if and only if the row and column permutations are equal.*

Lemma 4 (Proposition in [44])**.** *Suppose that $\hat{s}_{t,i}$ and $\hat{s}_{t,j}$ are conditionally independent given $\hat{\mathbf{z}}_t$. Then, for all $\hat{\mathbf{z}}_t$,*

$$\frac{\partial^2 \log p(\hat{\mathbf{s}}_t \mid \hat{\mathbf{z}}_t)}{\partial \hat{s}_{t,i} \partial \hat{s}_{t,j}} = 0.$$

Proof. Let $(\hat{\mathbf{z}}_t, \hat{\mathbf{s}}_t, \hat{g}_m)$ be the estimations of $(\mathbf{z}_t, \mathbf{s}_t, g_m)$. By Lemma 2,

$$\mathbf{x}_t = g_m(\mathbf{z}_t, \mathbf{s}_t); \quad \hat{\mathbf{x}}_t = \hat{g}_m(\hat{\mathbf{z}}_t, \hat{\mathbf{s}}_t) \tag{A33}$$

Suppose we reconstruct observations well: $\mathbf{x}_t = \hat{\mathbf{x}}_t$. Combined with Theorem 1,

$$p(\mathbf{x}_t \mid \hat{\mathbf{z}}_t) = p(\mathbf{x}_t \mid h_z(\mathbf{z}_t)) = p(\mathbf{x}_t \mid \mathbf{z}_t) \implies p(g_m(\mathbf{s}_t, \mathbf{z}_t) \mid \mathbf{z}_t) = p(\hat{g}_m(\hat{\mathbf{s}}_t, \hat{\mathbf{z}}_t) \mid \hat{\mathbf{z}}_t). \tag{A34}$$

Corollary A1 has shown that $\mathbf{J}_m(\mathbf{s}_t)$ and $\mathbf{J}_{\hat{g}_m}(\hat{\mathbf{s}}_t)$ are invertible matrices, by the definition of partial Jacobian matrix: $[\mathbf{J}_m(\mathbf{s}_t)]_{i,j} = \frac{\partial x_{t,i}}{\partial s_{t,j}} = \frac{\partial g_{m_i}(\mathbf{s}_t, \mathbf{z}_t)}{\partial s_{t,j}}$,

$$\frac{1}{|\mathbf{J}_m(\mathbf{s}_t)|} p(\mathbf{s}_t \mid \mathbf{z}_t) = \frac{1}{|\mathbf{J}_{\hat{g}_m}(\hat{\mathbf{s}}_t)|} p(\hat{\mathbf{s}}_t \mid \mathbf{z}_t). \tag{A35}$$

We define $h_s := m^{-1} \circ \hat{g}_m$ for any fixed $\mathbf{z}_t$ and $\hat{\mathbf{z}}_t$, hence, $|\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)| = \frac{|\mathbf{J}_{\hat{g}_m}(\hat{\mathbf{s}}_t)|}{|\mathbf{J}_m(\mathbf{s}_t)|}$ and $\hat{\mathbf{s}}_t = h_s(\mathbf{s}_t)$. Therefore, we have

$$p(\hat{\mathbf{s}}_t \mid \mathbf{z}_t) = \frac{1}{|\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)|} p(\mathbf{s}_t \mid \mathbf{z}_t) \implies \log p(\hat{\mathbf{s}}_t \mid \hat{\mathbf{z}}_t) = \log p(\mathbf{s}_t \mid \mathbf{z}_t) - \log |\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)|. \tag{A36}$$

The second-order partial derivative of $\log p(\hat{\mathbf{s}}_t \mid \hat{\mathbf{z}}_t)$ w.r.t. $(\hat{s}_{t,i}, \hat{s}_{t,j})$ is

$$\frac{\partial \log p(\hat{\mathbf{s}}_t \mid \hat{\mathbf{z}}_t)}{\partial \hat{s}_{t,i}} = \sum_{k=1}^{n} \frac{\partial \mathbf{A}_{t,k}}{\partial s_{t,k}} \cdot \frac{\partial s_{t,k}}{\partial \hat{s}_{t,i}} - \frac{\partial \log |\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)|}{\partial \hat{s}_{t,i}} = \sum_{k=1}^{n} \frac{\partial \mathbf{A}_{t,k}}{\partial s_{t,k}} \cdot [\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)]_{k,i} - \frac{\partial \log |\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)|}{\partial \hat{s}_{t,i}},$$
$$\frac{\partial^2 \log p(\hat{\mathbf{s}}_t \mid \hat{\mathbf{z}}_t)}{\partial \hat{s}_{t,i} \partial \hat{s}_{t,j}} = \sum_{k=1}^{n} \left(\frac{\partial^2 \mathbf{A}_{t,k}}{\partial s_{t,k}^2} \cdot [\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)]_{k,i} \cdot [\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)]_{k,j} + \frac{\partial \mathbf{A}_{t,k}}{\partial s_{t,k}} \cdot \frac{\partial [\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)]_{k,i}}{\partial \hat{s}_{t,j}} \right) - \frac{\partial^2 \log |\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)|}{\partial \hat{s}_{t,i} \partial \hat{s}_{t,j}}. \tag{A37}$$

Since for any $(i, j, t) \in \mathcal{J} \times \mathcal{J} \times \mathcal{T}$, we have $s_{t,i} \perp\!\!\!\perp s_{t,j} \mid \mathbf{z}_t$, Lemma 4 tells us $\frac{\partial^2 \log p(\hat{\mathbf{s}}_t|\hat{\mathbf{z}}_t)}{\partial \hat{s}_{t,i} \partial \hat{s}_{t,j}} = 0$. Therefore, its partial derivative w.r.t. $z_{t,l}$ ($l \in \mathcal{J}$) is always 0:

$$\frac{\partial^3 \log p(\hat{\mathbf{s}}_t \mid \hat{\mathbf{z}}_t)}{\partial \hat{s}_{t,i} \partial \hat{s}_{t,j} \partial z_{t,l}} = \sum_{k=1}^{n} \left(\frac{\partial^3 \mathbf{A}_{t,k}}{\partial s_{t,k}^2 \partial z_{t,l}} \cdot [\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)]_{k,i} \cdot [\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)]_{k,j} + \frac{\partial^2 \mathbf{A}_{t,k}}{\partial s_{t,k} \partial z_{t,l}} \cdot \frac{\partial [\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)]_{k,i}}{\partial \hat{s}_{t,j}} \right) \equiv 0, \tag{A38}$$

since entries of $\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)$ do not depend on $z_{t,l}$. By Assumption 3, maintaining this equality requires $[\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)]_{k,i} \cdot [\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)]_{k,j} = 0$ for $i \neq j$, which implies $\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t)$ is a monomial matrix.

Eliminate the Permutation Indeterminacy. We leverage the following properties:

1. The inverse of a lower triangular matrix remains a lower triangular matrix.
2. A matrix representing a DAG can always be permuted into a lower-triangular form using appropriate row and column permutations.
3. Corollary A2 states that:

$$\mathbf{J}_{g^L}(\mathbf{x}_t) = \mathbf{I}_{d_x} - \mathbf{D}_{m^L}(\mathbf{s}_t)\mathbf{J}_{m^L}^{-1}(\mathbf{s}_t); \quad \mathbf{J}_g(\mathbf{x}_t) = \mathbf{I}_{d_x} - \mathbf{D}_m(\mathbf{s}_t)\mathbf{J}_m^{-1}(\mathbf{s}_t) \tag{A39}$$

where $\mathbf{J}_{g^L}(\mathbf{x}_t)$ and $\mathbf{J}_{m^L}(\mathbf{s}_t)$ are (strictly) lower triangular matrices obtained by permuting $\mathbf{J}_g(\mathbf{x}_t)$ and $\mathbf{J}_m(\mathbf{s}_t)$, respectively. $\mathbf{D}_{m^L}(\mathbf{s}_t)$ is the diagonal matrix extracted from $\mathbf{J}_{m^L}(\mathbf{s}_t)$. Consequently, we can express the relationship between $\mathbf{J}_m(\mathbf{s}_t)$ and $\mathbf{J}_{m^L}(\mathbf{s}_t)$ as follows:

$$\mathbf{J}_{g^L}(\mathbf{x}_t) = \mathbf{P}_{d_x}\mathbf{J}_g(\mathbf{x}_t)\mathbf{P}_{d_x}^\top \implies \mathbf{J}_m(\mathbf{s}_t) = \mathbf{P}_{d_x}\mathbf{J}_{m^L}(\mathbf{s}_t)\mathbf{D}_{m^L}^{-1}(\mathbf{s}_t)\mathbf{P}_{d_x}^\top\mathbf{D}_m(\mathbf{s}_t), \tag{A40}$$

where $\mathbf{P}_{d_x}$ is the Jacobian matrix of a permutation function on the d_x-dimensional vector. Consequently, by $\mathbf{J}_m(\mathbf{s}_t) = \mathbf{J}_{\hat{g}_m}(\hat{\mathbf{s}}_t)\mathbf{J}_{h_s}(\mathbf{s}_t)$, we obtain

$$\mathbf{J}_{\hat{g}_m}(\hat{\mathbf{s}}_t) = \mathbf{P}_{d_x}\mathbf{J}_{m^L}(\mathbf{s}_t)\mathbf{D}_{m^L}^{-1}(\mathbf{s}_t)\mathbf{P}_{d_x}^\top\mathbf{D}_m(\mathbf{s}_t)\mathbf{J}_{h_s}^{-1}(\mathbf{s}_t), \tag{A41}$$

Using Lemma 3, we obtain $\mathbf{P}_{d_x}\mathbf{D}_{m^L}^{-1}(\mathbf{s}_t)\mathbf{P}_{d_x}^\top\mathbf{D}_m(\mathbf{s}_t)\mathbf{J}_{h_s}(\hat{\mathbf{s}}_t) = \mathbf{I}_{d_x}$, which implies $\mathbf{J}_{h_s}^{-1}(\mathbf{s}_t) = \mathbf{D}_m^{-1}(\mathbf{s}_t)\mathbf{D}_{m^L}(\mathbf{s}_t)$, a diagonal matrix. Consequently, $\mathbf{J}_{\hat{g}_m}(\hat{\mathbf{s}}_t)$ and $\mathbf{J}_m(\mathbf{s}_t)$ have the same support, meaning $\mathbf{J}_{\hat{g}}(\hat{\mathbf{x}}_t)$ and $\mathbf{J}_g(\mathbf{x}_t)$ share the same support as well, according to Corollary A2. Thus, by Assumption 2, the structure of the observational causal graph is identifiable. □

Discussion on Assumptions. To enhance understanding of our theoretical results, we provide some explanations of the assumptions, their connections to real-world scenarios, as well as the potential boundaries of theoretical results.

i. **Generation Variability.** Sufficient changes on generation 3 is widely used in identifiable nonlinear ICA/causal representation learning [27, 39, 32, 94, 86]. In practical climate science, it has been demonstrated that, within a given region, human activities ($s_{t,i}$) are strongly impacted by certain high-level climate latent variables $\mathbf{z}_t$ [1], following a process with sufficient changes [51].

ii. **Functional faithfulness.** Functional faithfulness corresponds to the *edge minimality* [91, 42, 60] for the Jacobian matrix $\mathbf{J}_g(\mathbf{x}_t)$ representing the nonlinear SEM $\mathbf{x}_t = g(\mathbf{x}_t, \mathbf{z}_t, \boldsymbol{\epsilon}_{\mathbf{x}_t})$, where $\frac{\partial x_{t,j}}{\partial x_{t,i}} = 0$ implies no causal edge, and $\frac{\partial x_{t,j}}{\partial x_{t,i}} \neq 0$ indicates causal relation $x_{t,i} \rightarrow x_{t,j}$. This assumption is fundamental to ensuring that the Jacobian matrix reflects the true causal graph. If our functional faithfulness is violated, the results can be misleading, but in theory (classical) faithfulness [75] is generally possible as discussed in [42] (2.3 Minimality). As a weaker version of it, edge minimality holds the same property. If needed, violations of faithfulness can be testable except in the triangle faithfulness situation [91]. As opposed to classical faithfulness [75], for example, this is not an assumption about the underlying world, but a convention to avoid redundant descriptions.

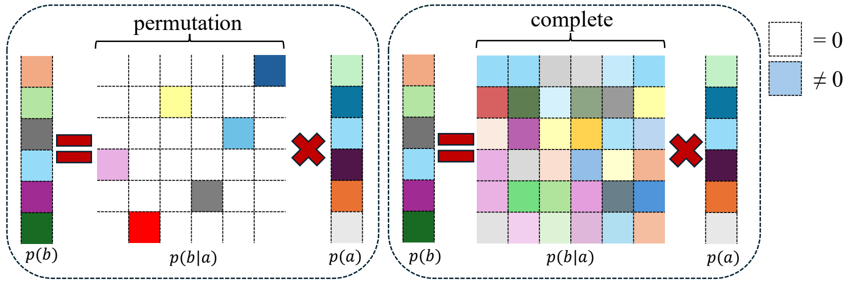


Figure A3: **Invertible Function v.s. Injective Operator.** *(Left)* Consider two variables a and b connected by the function $b = g(a)$, where g is invertible. *(Right)* Alternatively, their relationship can be expressed as $p(b) = L_{b|a} \circ p(a)$, where $L_{b|a}$ is an injective operator. The grid represents $p(b \mid a)$, with color indicating non-zero values and white representing zero. Intuitively, in the discrete case, a full-rank matrix corresponds to this relationship.

A.8 Proof of Lemma 1

(We delay the section of this proof since it relies on previous results.) The injectivity of a operator is formally characterized by the completeness of the conditional density function $p(a \mid b)$ used in the operator, as defined below.

Definition 5 (Completeness)**.** *A family of conditional density functions $p_{A|B}$ is said to be complete if the only solution to $\int_A p(a)p_{a|b}(a \mid b)\, da = 0,\ \forall b \in \mathcal{B}$ is $p(a) = 0$.*

Since the transformation from $\mathbf{s}_t$ to $\mathbf{x}_t$ is invertible and deterministic, given a $\dot{\mathbf{s}}_t \in \mathcal{S}_t$, the probability density function for $\mathbf{x}_t$ can be expressed as: $p(\mathbf{x}_t) = \begin{cases} \frac{1}{|\mathbf{J}_m(\mathbf{s}_t)|} p(\mathbf{s}_t), & \mathbf{x}_t = m(\mathbf{s}_t) \\ 0, & \mathbf{x}_t \neq m(\mathbf{s}_t) \end{cases}$. Hence, the conditional probability can be represented using the Dirac delta function:

$$p(\mathbf{x}_t \mid \mathbf{s}_t) = \delta(\mathbf{x}_t - m(\mathbf{s}_t)) \implies p(\mathbf{x}_t) = L_{\mathbf{x}_t|\mathbf{s}_t} \circ p(\mathbf{s}_t) = \int_{\mathcal{S}_t} \delta(\mathbf{x}_t - m(\mathbf{s}_t)) p(\mathbf{s}_t)\, d\mathbf{s}_t.$$

By recalling Eq. (2), we can rewrite $p(\mathbf{x}_t)$ in terms of the operator $L_{\mathbf{x}_t|\mathbf{s}_t}$ acting on $p_{\mathbf{s}_t}$. We consider $p(\mathbf{x}_t \mid \mathbf{s}_t)$ as an infinite-dimensional vector, and the operator $L_{\mathbf{x}_t|\mathbf{s}_t}$ as an infinite-dimensional matrix where

$$L_{\mathbf{x}_t|\mathbf{s}_t} = [\delta(\mathbf{x}_t - m(\mathbf{s}_t))]^{\top}_{\mathbf{x}_t \in \mathcal{X}_t}.$$

By Corollary A2, since $\mathbf{J}_m(\mathbf{s}_t)$ is invertible, for any two different points $\mathbf{s}_t^{(1)}, \mathbf{s}_t^{(2)} \in \mathcal{S}_t$ ($\mathbf{s}_t \neq \mathbf{s}'_t$), we have $m(\mathbf{s}_t^{(1)}) \neq m(\mathbf{s}_t^{(2)})$. This implies that the supports of $\delta(\mathbf{x}_t - m(\mathbf{s}_t^{(1)}))$ and $\delta(\mathbf{x}_t - m(\mathbf{s}_t^{(2)}))$ are disjoint. Thus, $[\delta(\mathbf{x}_t - m(\mathbf{s}_t))]^{\top}_{\mathbf{x}_t \in \mathcal{X}_t}$ preserves a one-to-one correspondence across the $\mathcal{X}_t$, ensuring:

$$\text{null}\ [\delta(\mathbf{x}_t - m(\mathbf{s}_t))]^{\top}_{\mathbf{x}_t \in \mathcal{X}_t} = \{0^{(\infty)}\},$$

which denotes the completeness of $L_{\mathbf{x}_t|\mathbf{s}_t}$ stated in Definition 5, indicating that $L_{\mathbf{x}_t|\mathbf{s}_t}$ is injective. The visualization in Figure A3 highlights why Assumption 1 is significantly less restrictive than the invertibility assumption adopted in most of the previous CRL literature [28, 29, 31, 34, 94, 43].

A.9 Comparison with Existing Methods

Our method targets the joint identification of latent causal graphs and observational causal structures in time series. This is essential for domains such as climate science, where latent processes govern observed dynamics. In contrast, IDOL [43] focuses solely on recovering latent variables and assumes a deterministic mixing function without any causal relations among observed variables. As a result, it cannot recover the observational causal graph and fails in contexts like climate systems, where both latent and observational structures are crucial.

Prior works [54, 64] use nonlinear ICA to recover causal relations among observed variables, assuming known domain variables and non-i.i.d. data. However, (1) CaDRe does not require predefined domain variables and instead leverages contextual information to infer latent variables as conditional priors; (2) ICA-based methods are not robust under latent confounding, as spurious correlations may obscure true causal links; (3) they do not identify the latent variables or their underlying dynamics; (4) they require invertibility of g and m, an assumption we relax in Corollary A1.

The FCI algorithm [74] allows for latent confounding and uses conditional independence tests to infer causal relations among observed variables. However, it cannot recover the latent variables themselves or their causal influence on observations. Furthermore, the causal structure is expressed as a Partial Ancestral Graph (PAG), which represents an equivalence class and may contain ambiguous or uncertain edges. Such ambiguity is particularly problematic for applications like climate analysis, which demand interpretable and stable causal structures.

B Related Work

B.1 Climate Analysis

Climate analysis is learning to address the complex, nonlinear, and high-dimensional nature of Earth system dynamics. A prominent line of work focuses on using neural networks for weather and climate forecasting, including data-driven models such as FourCastNet [58] and GraphCast [41], which demonstrate remarkable predictive performance by modeling spatiotemporal dependencies. However, these methods often lack interpretability and fail to reveal the underlying causal mechanisms driving climate variability. To address this issue, recent research has integrated causal discovery into climate science. For instance, [68] introduces causal inference frameworks tailored to climate time series, incorporating techniques such as PCMCI to infer lagged and contemporaneous dependencies. Other approaches employ structural causal models (SCMs) for identifying interactions between climate variables under interventions [63]. Beyond shallow models, efforts have emerged to disentangle latent variables in high-dimensional climate data using variational autoencoders [35]. While effective, most of these methods do not guarantee identifiability or robust generalization across regimes.

More recently, hybrid models that couple dynamical systems theory with deep learning have shown promise in capturing climate processes with greater fidelity. Examples include integrating physics-based constraints into latent state-space models [4] and learning interpretable representations for climate variability modes such as the Madden-Julian Oscillation [77]. These works highlight the growing interest in combining structure learning, causal inference, and deep latent modeling to move beyond black-box predictions towards actionable scientific understanding.

B.2 Causal Representation Learning

Achieving causal representations for time series data [61] often relies on nonlinear ICA to recover latent variables with identifiability guarantees [85, 71]. Classical ICA methods assume a linear mixing function between latent and observed variables [10]. To move beyond this linearity assumption, recent advances in nonlinear ICA have established identifiability under various alternative assumptions, including the use of auxiliary variables or structural sparsity [97, 30, 27]. One prominent line of work introduces auxiliary variables to facilitate identifiability. For instance, [32] achieves identifiability by assuming latent sources follow an exponential family distribution and incorporating side information such as domain, time, or class labels [28, 29, 31]. To relax the exponential family requirement, [36] establishes component-wise identifiability using $2n + 1$ auxiliary variables for n latent components. Another direction pursues identifiability in a fully unsupervised setting by leveraging structural sparsity. [39] propose a sparsity-based inductive bias to disentangle latent causal factors, demonstrating identifiability in multi-task learning and related settings. They further extend these results to establish identifiability up to a consistency class [38], allowing partial disentanglement. Complementarily, [97] and [94, 43] exploit sparse latent structures under distributional shifts to obtain identifiability results without relying on auxiliary information.

B.3 Causal Discovery

Existing causal discovery methods in climate analysis primarily build on extensions of PCMCI [69], which effectively captures time-lagged and instantaneous linear dependencies, and its nonlinear variant [67]. However, both approaches assume fully observed systems and neglect latent variables, limiting their applicability to complex climate dynamics. Recent causal representation learning methods motivated by climate science attempt to address this gap: [6] imposes strong identifiability assumptions via single-node structures, while [84] adopts an ODE-based model to study climate-zone classification, though these methods often overlook dependencies among observed variables. Beyond climate, a class of nonlinear causal discovery methods leverages Jacobian information for identifiability and acyclicity [37, 65], including applications to structural equation models [2], Markov structures [96], independent mechanisms [18], and non-i.i.d. settings [64]. While [12] propose a general framework that accounts for hidden variables by using rank conditions on the observed covariance matrix, their model is restricted to linear relationships and cannot recover nonlinear latent dynamics in time-series data. In contrast, our method, CaDRe, recovers latent causal structures under nonlinear dependencies, though it currently does not support cases where observed variables act as causes of latent ones—a limitation we leave to future work. Considering the nonlinear CD based on continuous optimization, we additionally provide the Table A2 for comparison.

B.4 Time-Series Forecasting

Time series forecasting has seen rapid progress with deep learning methods that leverage various neural architectures. RNN-based models [22, 40, 70] focus on sequential dependencies, while CNN-based approaches [3, 79, 81] capture local temporal patterns. State-space models [20, 19, 21] offer structured modeling of latent dynamics. Transformer-based methods [99, 82, 57] further advance long-range forecasting through attention mechanisms. However, most existing methods neglect instantaneous dependencies among variables, limiting their ability to fully capture the joint dynamics of multivariate time series.

Table A2: Comparison of different methods based on their properties in function type (f), data, Jacobian (J), capability of performing Causal Discovery (CD) and Causal Representation Learning (CRL), and whether they achieve identifiability.

Method	f	Data	J	CD	CRL	Identifiability
LiNGAM [72]	Linear	Non-Gaussian	$J_{f^{-1}}$	✔	✗	✔
GraN-DAG [37]	Additive	Gaussian	$J_{f^{-1}}$	✔	✗	✗
IMA [18]	IMA	All	J_f	✗	✗	✔
G-SCM [96]	Sparse	All	J_f	✗	✗	✔
Score-Based FCMs [65]	Additive	Gaussian	$J_{\nabla_x \log p(x)}$	✔	✗	✗
DynGFN [2]	Cyclic (ODE)	All	J_f	✔	✗	✗
JCD [64]	All	Assums. 2, F. 1	$J_{f^{-1}}$	✔	✗	✔
CausalScore [47]	Mixed	Gaussian	$J_{\nabla_x \log p(x)}$	✔	✗	Partial
CaDRe (Ours)	All	All	$J_{f^{-1}}$	✔	✔	✔

C Estimation Methodology

Our theoretical insights shed light on the practical implementations. As shown in Figure A5, we instantiate these insights into an estimation framework for **Ca**usal **D**iscovery and causal **Re**presentation learning (**CaDRe**) in the nonparametric setting, enabling direct inference of causal structures.

Overall Architecture. The proposed architecture is built upon the variational autoencoder [33]. In light of data generating process 1, we establish the Evidence Lower BOund (ELBO) as follows:

$$\begin{aligned}\mathcal{L}_{ELBO} = \mathbb{E}_{q(\mathbf{s}_{1:T}|\mathbf{x}_{1:T})}\left[\log p(\mathbf{x}_{1:T} \mid \mathbf{s}_{1:T}, \mathbf{z}_{1:T})\right] - \\ \lambda_1 D_{\text{KL}}\left(q(\mathbf{s}_{1:T} \mid \mathbf{x}_{1:T}) \,\|\, p(\mathbf{s}_{1:T} \mid \mathbf{z}_{1:T})\right) - \lambda_2 D_{\text{KL}}\left(q(\mathbf{z}_{1:T} \mid \mathbf{x}_{1:T}) \,\|\, p(\mathbf{z}_{1:T})\right),\end{aligned} \tag{A42}$$

where λ_1 and λ_2 are hyperparameters, and D_{KL} represents the Kullback-Leibler divergence. We set $\lambda_1 = 4 \times 10^{-3}$ and $\lambda_2 = 1.0 \times 10^{-2}$ to achieve the best performance. In Figure A5, the `z-encoder`, `s-encoder` and `decoder` are implemented by Multi-Layer Perceptrons (MLPs) as follows:

$$\mathbf{z}_{1:T} = \phi(\mathbf{x}_{1:T}),\ \mathbf{s}_{1:T} = \eta(\mathbf{x}_{1:T}),\ \hat{\mathbf{x}}_{1:T} = \psi(\mathbf{z}_{1:T}, \mathbf{s}_{1:T}), \tag{A43}$$

respectively, where the `z-encoder` ϕ learns the latent variables through denoising, and `s-encoder` ψ and `decoder` η approximate functions for encoding $\mathbf{s}_t$ and reconstructing observations, respectively.
Prior Estimation of $\mathbf{z}_t$ and $\mathbf{s}_t$. We propose using the `s-prior` network and `z-prior` network to recover the independent noise $\hat{\boldsymbol{\epsilon}}_t^x$ and $\hat{\boldsymbol{\epsilon}}_t^z$, respectively, thereby estimating the prior distribution of latent variables $\hat{\mathbf{z}}_t$ and dependent noise $\hat{\mathbf{s}}_t$. Specifically, we first let r_i be the i-th learned inverse transition function that takes the estimated latent variables as input to recover the noise term, *e.g.*, $\hat{\epsilon}_{t,i}^z = r_i(\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_t)$. Each r_i is implemented by MLPs. Sequentially, we devise a transformation $\kappa := \{\hat{\mathbf{z}}_{t-1}, \hat{\mathbf{z}}_t\} \to \{\hat{\mathbf{z}}_{t-1}, \hat{\boldsymbol{\epsilon}}_t^z\}$, whose Jacobian can be formalized as $\mathbf{J}_\kappa = \begin{pmatrix} \mathbf{I} & 0 \\ \mathbf{J}_d(\hat{\mathbf{z}}_{t-1}) & \mathbf{J}_r(\hat{\mathbf{z}}_t) \end{pmatrix}$.

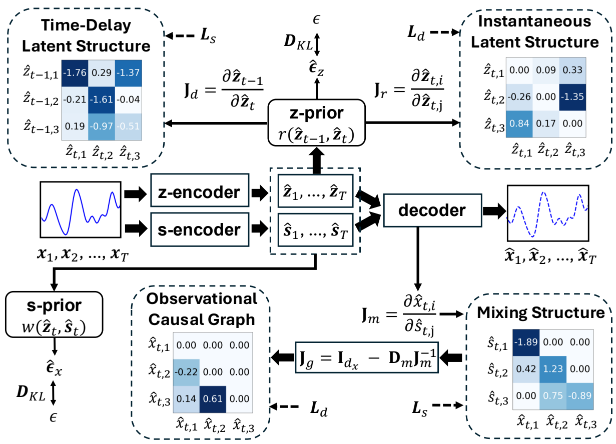


Figure A4: **The estimation procedure of CaDRe.** The model framework includes two encoders: `z-encoder` for extracting latent variables $\mathbf{z}_t$, and `s-encoder` for extracting $\mathbf{s}_t$. A `decoder` reconstructs observations from these variables. Additionally, prior networks estimate the prior distribution using normalizing flow, target on learning causal structure based on the Jacobian matrix. $\mathcal{L}s$ imposes a sparsity constraint and $\mathcal{L}d$ enforces the DAG structure on Jacobian matrix. D_{KL} enforces an independence constraint on the estimated noise by minimizing its KL divergence w.r.t. $\mathcal{N}(0, \mathbf{I})$. In summary, this method learns independent noise to inversely infer the causal structures.

Then we have Eq. (A44) derived from normalizing flow to estimate the prior distribution:

$$\log p(\hat{\mathbf{z}}_t, \hat{\mathbf{z}}_{t-1}) = \log p(\hat{\mathbf{z}}_{t-1}, \hat{\boldsymbol{\epsilon}}_t^z) + \log |\frac{\partial r_i}{\partial \hat{z}_{t,i}}|. \tag{A44}$$

According to the generation process, the noise $\epsilon_{t,i}^z$ is independent of $\mathbf{z}_{t-1}$, allowing us to enforce independence on the estimated noise term $\hat{\epsilon}_{t,i}^z$ with $\mathcal{D}_{KL}$. Consequently, Eq. (A44) can be rewritten as:

$$\log p(\hat{\mathbf{z}}_{1:T}) = p(\hat{\mathbf{z}}_1) \prod_{\tau=2}^{T} \left(\sum_{i=1}^{d_z} \log p(\hat{\epsilon}_{\tau,i}^z) + \sum_{i=1}^{d_z} \log |\frac{\partial r_i}{\partial \hat{z}_{\tau,i}}| \right), \tag{A45}$$

where $p(\hat{\epsilon}_{\tau,i}^z)$ is assumed to follow a Gaussian distribution. Similarly, we estimate the prior of $\mathbf{s}_t$ using $\hat{\epsilon}_{t,i}^x = w_i(\hat{\mathbf{z}}_t, \hat{\mathbf{s}}_t)$, and model the transformation between $\hat{\mathbf{s}}_t$ and $\hat{\mathbf{z}}_t$ as follows:

$$\log p\left(\hat{\mathbf{s}}_{1:T} \mid \hat{\mathbf{z}}_{1:T}\right) = \prod_{\tau=1}^{T} \left(\sum_{i=1}^{d_x} \log p\left(\hat{\epsilon}_{\tau,i}^x\right) + \sum_{i=1}^{d_x} \log \left| \frac{\partial w_i}{\partial \hat{s}_{\tau,i}} \right| \right). \tag{A46}$$

Specifically, to ensure the conditional independence of components in $\hat{\mathbf{z}}_t$ and $\hat{\mathbf{s}}_t$, we using $\mathcal{D}_{KL}$ to minimize the KL divergence from the distributions of $\hat{\boldsymbol{\epsilon}}_t^x$ and $\hat{\boldsymbol{\epsilon}}_t^z$ to the distribution $\mathcal{N}(0, \mathbf{I})$.

Structure Learning. The variables r_i and w_i are designed to capture causal dependencies among latent and observed variables, respectively. We denote $\mathbf{J}_d(\hat{\mathbf{z}}_{t-1})$ as the Jacobian matrix of the function r, which implies the estimated time-lagged latent causal structure; $\mathbf{J}_r(\hat{\mathbf{z}}_t)$, which implies the estimation of instantaneous latent causal structure; and $\mathbf{J}_{\hat{g}}(\hat{\mathbf{x}}_t)$, which implies the estimated observational causal graph. Considering the observational causal graph, we compute $\mathbf{J}_{\hat{m}}(\hat{\mathbf{s}}_t)$ from the `decoder`, and instantly obtain the observational causal graph $\mathbf{J}_{\hat{g}}(\hat{\mathbf{x}}_t)$ via Corollary A2. Notably, the entries of $\mathbf{J}_{\hat{g}}(\hat{\mathbf{x}}_t)$ vary with other variables such as $\hat{\mathbf{z}}_t$, resulting in a DAG that could change over time. For the latent structure, we directly compute $\mathbf{J}_d(\hat{\mathbf{z}}_{t-1})$ and $\mathbf{J}_r(\hat{\mathbf{z}}_t)$ from `z-prior` network as the time-lagged structure and instantaneous structure in latent space, respectively. To prevent redundant edges and cycles, a sparsity penalty $\mathcal{L}_s$ are imposed on each learned structure, and DAG constraints $\mathcal{L}_d$ are imposed on the observational causal graph and instantaneous latent causal DAG. Specifically, the Markov network structure for latent variables is derived as $\mathcal{M}(\mathbf{J}) = (\mathbf{I} + \mathbf{J})^\top (\mathbf{I} + \mathbf{J}) - \mathbf{I}$. Formally, we define these penalties as follows:

$$\sum \mathcal{L}_s = \|\mathcal{M}(\mathbf{J}_r(\hat{\mathbf{z}}_t))\|_1 + \|\mathcal{M}(\mathbf{J}_d(\hat{\mathbf{z}}_{t-1}))\|_1 + \|\mathbf{J}_{\hat{g}}(\hat{\mathbf{x}}_t)\|_1, \quad \sum \mathcal{L}_d = \mathcal{D}(\mathbf{J}_{\hat{g}}(\hat{\mathbf{x}}_t)) + \mathcal{D}(\mathbf{J}_r(\hat{\mathbf{z}}_t)). \tag{A47}$$

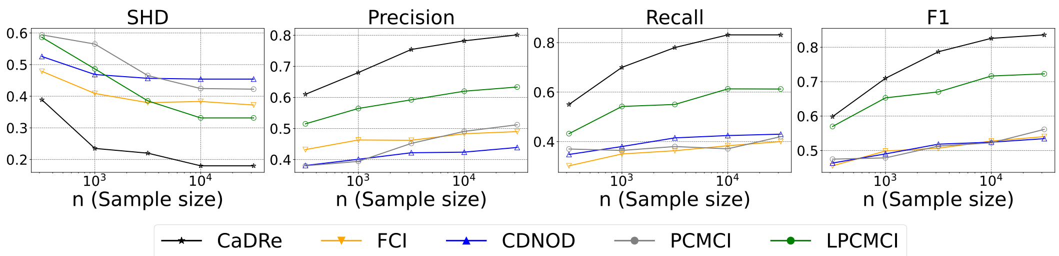
where $\mathcal{D}(A) = \operatorname{tr}\left[(I + \frac{1}{d} A \circ A)^d\right] - d$ is the DAG constraint from [89], with A being an d-dimensionality matrix. $||\cdot||_1$ denotes the matrix l_1 norm. In summary, the overall loss function of the CaDRe model integrates ELBO and penalties for structural constraints, which is formalized as:

$$\mathcal{L}_{ALL} = \mathcal{L}_{ELBO} + \alpha \sum \mathcal{L}_s + \beta \sum \mathcal{L}_d, \tag{A48}$$

Table A3: **Results under Varying Observational Dimensionality (d_x).** Each setting is repeated with 5 random seeds. For evaluation, the best-converged result per seed is selected to avoid local minima.

d_z	d_x	SHD ($\mathbf{J}_{\hat{g}}(\hat{\mathbf{x}}_t)$)	TPR	Precision	MCC ($\mathbf{s}_t$)	MCC ($\mathbf{z}_t$)	SHD ($\mathbf{J}_r(\hat{\mathbf{z}}_t)$)	SHD ($\mathbf{J}_d(\hat{\mathbf{z}}_{t-1})$)	R^2
3	3	0	1	1	0.9775$\pm$ 0.01	0.9721$\pm$ 0.01	0.27$\pm$ 0.05	0.26$\pm$ 0.03	0.90$\pm$ 0.05
	6	0.18$\pm$ 0.06	0.83$\pm$ 0.03	0.80$\pm$ 0.04	0.9583$\pm$ 0.02	0.9505$\pm$ 0.01	0.24$\pm$ 0.06	0.33$\pm$ 0.09	0.92$\pm$ 0.01
	8	0.29$\pm$ 0.05	0.78$\pm$ 0.05	0.76$\pm$ 0.04	0.9020$\pm$ 0.03	0.9601$\pm$ 0.03	0.36$\pm$ 0.11	0.31$\pm$ 0.12	0.93$\pm$ 0.02
	10	0.43$\pm$ 0.05	0.65$\pm$ 0.08	0.63$\pm$ 0.14	0.8504$\pm$ 0.07	0.9652$\pm$ 0.02	0.29$\pm$ 0.04	0.40$\pm$ 0.05	0.92$\pm$ 0.02
	100*	0.17$\pm$ 0.02	0.80$\pm$ 0.05	0.81$\pm$ 0.02	0.9131$\pm$ 0.02	0.9565$\pm$ 0.02	0.21$\pm$ 0.01	0.29$\pm$ 0.10	0.93$\pm$ 0.03

Figure A6: **Comparison with Constraint-Based CD.** We set $d_x = 6$ and $d_z = 3$. We run experiments using 5 different random seeds, and report the average performance on evaluation metrics.



where $\alpha = 1.0 \times 10^{-4}$ and $\beta = 5.0 \times 10^{-5}$ are hyperparameters. The discussions about hyperparameter selections and their effects on performance are given in Appendix E.2.

D Experimental Results

Based on the proposed framework, we conduct extensive experiments on both synthetic and real-world climate data to examine the identifiability of the latent process and observational causal graph, as well as climate forecasting and scientific interpretability in realistic climate systems.

D.1 On Synthetic Climate Data

Baselines. The data simulation processes and evaluation metrices are presented in Appendix E.2. In CD, we compare CaDRe with several constraint-based methods suited for nonparametric settings. Specifically, we include FCI [73] and CD-NOD [26], which handle latent confounders, and time-series methods PCMCI [69] and LPCMCI [17], which account for instantaneous and lagged effects with latent confounding. In CRL, we benchmark against CaRiNG [8], TDRL [86], LEAP [87], SlowVAE [34], PCL [29], i-VAE [32], TCL [28], and models that handle instantaneous effects, including iCITRIS [45] and G-CaRL [55]. Details are presented in Appendix E.2.

D.2 On Real-World Climate Data

Baselines. Details about the climate datasets are presented in Appendix E.3. We consider the following state-of-the-art deep forecasting models for time series forecasting. First, we consider the conventional methods for time series forecasting, including Autoformer [82], TimesNet [81] and MICN [79]. Moreover, we consider several latest methods for time series analysis like CARD [80], FITS [83], and iTransformer [48]. Finally, we consider the TDRL [86]. We repeat each experiment over 3 random seeds and publish the average performance.

Causal Discovery Consistency. As the ground-truth causal graph is inaccessible in real climate data, we adopt the contemporaneous wind field [62] as a surrogate for evaluation. As shown in Figure A7, CaDRe recovers observational causal graphs closely consistent with physical wind patterns, serving as a scientific support. Specifically, CaDRe captures large-scale physical patterns (*e.g.*, westward flows in equatorial oceans, southwestward propagation near Central America), while revealing structurally complex zones along coastal boundaries. These dense, irregular edges may reflect coupled land–atmosphere dynamics or anthropogenic influences [78, 5]. The latent transition $\hat{\mathbf{z}}_{t-1} \rightarrow \hat{\mathbf{z}}_t$ is also visualized to unveil the hidden dynamic process in the scientific discovery.

Weather Prediction. We evaluate our method on the CESM2 sea surface temperature dataset for real-world temperature forecasting. As summarized in Table A5, our approach outperforms existing

Table A4: **Identification Results on Simulated Data.** We set the dimensions as $d_z = 3$ and $d_x = 10$, and consider three scenarios according to our theory: *i) Independent*: $z_{t,i}$ and $z_{t,j}$ are conditionally independent given $\mathbf{z}_{t-1}$; *ii) Sparse*: $z_{t,i}$ and $z_{t,j}$ are dependent given $\mathbf{z}_{t-1}$, but the latent Markov network $\mathcal{G}_{z_t}$ and time-lagged latent structure are sparse; *iii) Dense*: No sparsity restrictions on latent causal graph. Bold numbers indicate the best performance.

Setting	Metric	CaDRe	iCITRIS	G-CaRL	CaRiNG	TDRL	LEAP	SlowVAE	PCL	i-VAE	TCL
Independent	MCC	**0.9811**	0.6649	0.8023	0.8543	0.9106	0.8942	0.4312	0.6507	0.6738	0.5916
	R^2	**0.9626**	0.7341	0.9012	0.8355	0.8649	0.7795	0.4270	0.4528	0.5917	0.3516
Sparse	MCC	**0.9306**	0.4531	0.7701	0.4924	0.6628	0.6453	0.3675	0.5275	0.4561	0.2629
	R^2	**0.9102**	0.6326	0.5443	0.2897	0.6953	0.4637	0.2781	0.1852	0.2119	0.3028
Dense	MCC	**0.6750**	0.3274	0.6714	0.4893	0.3547	0.5842	0.1196	0.3865	0.2647	0.1324
	R^2	**0.9204**	0.6875	0.8032	0.4925	0.7809	0.7723	0.5485	0.6302	0.1525	0.2060

Table A5: **Results on Temperature Forecasting.** Lower MSE/MAE is better. **Bold** numbers represent the best performance among the models, while <u>underlined</u> numbers denote the second-best.

Dataset	Predicted Length	CaDRe		TDRL		CARD		FITS		MICN		iTransformer		TimesNet		Autoformer	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
CESM2	96	<u>0.410</u>	<u>0.483</u>	0.439	0.507	**0.409**	**0.484**	0.439	0.508	0.417	0.486	0.422	0.491	0.415	0.486	0.959	0.735
	192	**0.412**	**0.487**	0.440	0.508	<u>0.422</u>	<u>0.493</u>	0.447	0.515	1.559	0.984	0.425	0.495	0.417	0.497	1.574	0.972
	336	**0.413**	**0.485**	0.441	0.505	<u>0.421</u>	<u>0.497</u>	0.482	0.536	2.091	1.173	0.426	0.494	0.423	0.499	1.845	1.078

time-series forecasting models in precision, due to existing models struggling with causally-related observations and non-contaminated generation, restricting their usability in real-world climate data.

E Experiment Details

E.1 Experiment Results on Simulated Datasets

Empirical Study. We show performance on the CD and CRL in Table A3, and investigate different dimensionalities of observed variables. Our results on both latent representation learning metrics verify the effectiveness of our methodology under identifiabilty, and the result on $d_x = 100$ makes it scalable to high-dimensional data, if prior knowledge of the elimination of some dependences are provided by the physical law of climate [14] or LLM [50], supports our subsequent experiment on real-world data. Additionally, the study on different d_z can be found in Appendix E.2.

Comparison with Constraint-Based CD. Figure A6 shows that CaDRe consistently outperforms all baselines across varying sample sizes, with performance improving as more data becomes available. In contrast, FCI performs poorly when latent confounders are dependent, often leading to low recall. CD-NOD relies on pseudo-causal sufficiency, assuming that latent variables are functions of surrogate variables, which does not hold in general latent settings. PCMCI ignores latent dynamics altogether, while LPCMCI assumes no causal relations among latent confounders, limiting its applicability in complex systems. These comparisons highlight the effectiveness of CaDRe in addressing the limitations of existing constraint-based methods.

Comparison with Temporal CRL. The MCC and R^2 results for the *independent* and *sparse* settings demonstrate that our model achieves component-wise identifiability (Theorem A.3). In contrast, other considered methods fail to recover latent variables, as they cannot properly address cases where the observed variables are causally-related. For the *dense* setting, our approach achieves monoblock identifiability (Theorem 1) with the highest R^2, while other methods exhibit significant degradation because they are not specifically tailored to handle scenarios involving general noise in the generating function. These outcomes are consistent with our theoretical analysis.

E.2 On Simulation Dataset

Data Simulation. We generate time series data with latent variables $\mathbf{z}_t \in \mathbb{R}^{d_z}$ and observed variables $\mathbf{x}_t \in \mathbb{R}^{d_x}$, where $d_z \leq d_x$. The latent dynamics follow a leaky non-linear autoregressive model:

$$\mathbf{z}_t = \sigma\left(\sum_{\ell=1}^{L} \mathbf{W}^{(\ell)} \mathbf{z}_{t-\ell}\right) + \boldsymbol{\epsilon}_t^z, \quad \boldsymbol{\epsilon}_t^z \sim \mathcal{N}(0, \sigma_z^2 \mathbf{I}), \tag{A49}$$

where $\sigma(\cdot)$ is leaky ReLU, and $\mathbf{W}^{(\ell)}$ are lag-ℓ transition matrices modulated by class-specific parameters. Instantaneous causal relations among $\mathbf{x}_t$ are defined by an Erdős-Rényi DAG $\mathbf{B} \in$

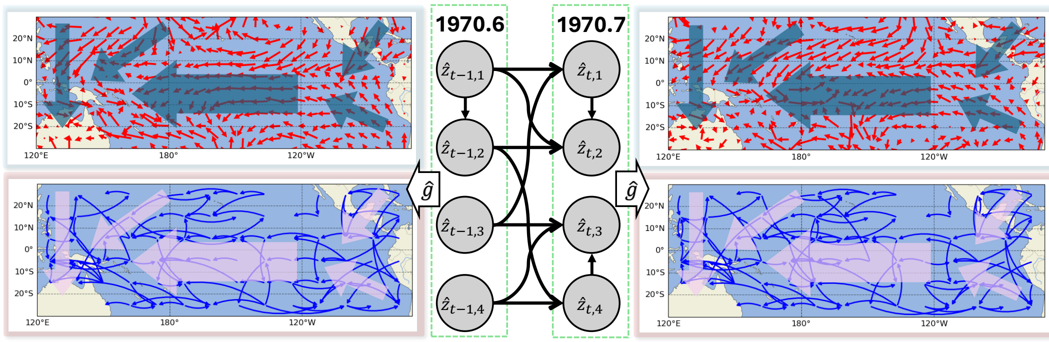


Figure A7: **Top:** Estimated instantaneous causal graph over climate grids. **Bottom:** Reference wind field from [62]. Blue arrows denote learned causal directions; red arrows indicate wind vectors.

$\{0,1\}^{d_x \times d_x}$, with time-varying edge weights:

$$\mathbf{B}_t = \alpha(t) \cdot \mathbf{B}, \quad \alpha(t) = a_1 \cos\left(\frac{2\pi t}{T}\right) + a_2. \tag{A50}$$

The observed variable $\mathbf{x}_t$ is first generated by a multilayer mixing of $(\mathbf{z}_t, \mathbf{s}_t)$, followed by additive and autoregressive noise:

$$\mathbf{x}_t = f_{\text{mix}}(\mathbf{z}_t, \mathbf{s}_t) + \mathbf{s}_t, \quad \mathbf{s}_t = \boldsymbol{\epsilon}_t^x + f_{\text{dep}}(\mathbf{x}_{t-1}), \tag{A51}$$

where $\boldsymbol{\epsilon}_t^x \sim \mathcal{U}[0, \sigma_x]$. Then, causal effects among observed variables are injected based on $\mathbf{B}_t$ in topological order:

$$x_{t,i} \leftarrow x_{t,i} + \sum_{j \in \text{pa}(i)} B_{t,j,i} \cdot x_{t,j}, \tag{A52}$$

where $\mathbf{pa}(i) = \{j \mid B_{t,j,i} \neq 0\}$ are the causal parents of variable i under $\mathbf{B}_t$.

Various Datasets. We generate simulated time-series data using the fixed latent causal process described in Eq. (1) and illustrated in Figure 1. To comprehensively evaluate our theoretical results, we construct synthetic datasets with varying observed dimensionalities, including $d_x = 3, 6, 8, 10, 100^{*}$[1] and latent dimensionalities $d_z = 2, 3, 4$, specified for each experiment. Additionally, we simulate different levels of structural sparsity in the latent process under three regimes: *Independent*, *Sparse*, and *Dense*. For evaluation, we use SHD, TPR, Precision, and Recall for causal structure recovery, and MCC and R^2 for assessing latent representation identifiability. As defined in Eq. (1), under the *Independent* setting for the latent temporal process and dependent noise variable $\mathbf{s}_t$, we use the generation process from [86], meaning there are no instantaneous dependencies within the $\mathbf{z}_t$. For *Sparse* and *Dense* settings, we gradually increase the graph degree after removing diagonals. Each independent noise is sampled from normal distributions.

Evaluation Metrics. We evaluate the recovery of latent variables and causal structures using the following metrics:

i. **Latent Space Recovery.** Following the identifiability result in Theorem 1, we measure the alignment between the estimated latent variables $\hat{\mathbf{z}}_t$ and the true latent variables $\mathbf{z}_t$ using the coefficient of determination R^2, where $R^2 = 1$ indicates perfect alignment. A nonlinear mapping is estimated using kernel regression with a Gaussian kernel.

ii. **Latent Component Recovery.** To evaluate component-wise identifiability as discussed in Theorem A1, we use the Spearman Mean Correlation Coefficient (MCC), which assesses the monotonic relationship between estimated and true latent components.

iii. **Latent Causal Structure.** For evaluating the recovery of latent causal graphs, both instantaneous and time-lagged, we compute the Structural Hamming Distance (SHD) between the learned and true adjacency matrices. Given the permutation indeterminacy of latent variables, we align the estimated latent causal structures $\mathbf{J}_r(\hat{\mathbf{z}}_t)$ and $\mathbf{J}_r(\hat{\mathbf{z}}_{t-1})$ with the ground truth by applying consistent permutations.

[1]* indicates the use of a masking scheme simulated from geographical information (see Appendix E.2)

iv. **Observational Source Recovery.** As a surrogate for evaluating the observational causal graph, we use MCC [32] to assess the recovery of $\mathbf{s}_t$. Unlike latent variables, this metric does not allow permutations and reflects the identifiability condition stated in Theorem 3.

v. **Causal Structure Accuracy.** The recovered latent and observational causal DAGs are also evaluated using SHD, normalized by the total number of possible edges to facilitate comparison across different graph sizes.

vi. **Graph-Level Metrics.** In addition to SHD, we report true positive rate (TPR), precision, and F1 score to benchmark our method against constraint-based approaches in causal graph recovery.

Implementation Details of CRL Baselines. We employed publicly available implementations for TDRL, CaRiNG, and iCRITIS, which cover most of the baselines used in our experiments. For G-CaRL, whose official code was not released, we re-implemented the method based on the descriptions in the original paper. Furthermore, because the original iCRITIS framework was tailored for image-based inputs, we adapted it to our setting by replacing its encoder and decoder with a VAE architecture, using the same hyperparameters as in CaDRe.

Mask by Inductive Bias. Continuous optimization faces challenges like local minima [56, 52], making it difficult to scale to higher dimensions. However, incorporating prior knowledge on the low probability of certain dependencies [75, 69] enables us to compute a mask. To validate this approach using physical laws as observed DAG initialization E.3 in climate data, we mask 75% of the lower triangular elements in a simulation with $d_x = 100$, a ratio much lower than in real-world applications.

Comparison with Constraint-Based Methods. We compare our method against a series of constraint-based causal discovery algorithms, which rely on Conditional Independence (CI) tests without assuming a specific form for the SEMs. These approaches are nonparametric and model-agnostic, but they typically return equivalence classes of graphs rather than fully identifiable structures. For instance, FCI outputs Partial Ancestral Graphs (PAGs), while CD-NOD returns equivalence classes reflecting causal ambiguity under the observed CI constraints. For a fair comparison, we adopt near-optimal configurations of the most representative constraint-based methods. Specifically, we use the `Causal-learn` package [95] to implement FCI and CD-NOD, and the `Tigramite` library [69] for PCMCI and LPCMCI. Each method is run under recommended hyperparameter settings as reported in their respective documentation or prior studies, ensuring a reliable and balanced comparison.

i. **FCI**: We use Fisher's Z conditional independence test. For the obtained PAG, we enumerate all possible adjacency matrices and select the one closest to the ground truth by minimizing the SHD.

ii. **CD-NOD**: We concatenate the time indices $[1, 2, \dots, T]$ of the simulated data into the observed variables and only consider the edges that exclude the time index. We use kernel-based CI test since it demonstrates superior performance here. We consider all obtained equivalence classes and select the result that minimizes SHD relative to the ground truth.

iii. **PCMCI**: We use partial correlation as the metric of the conditional independence test. We enforce no time-lagged relationships in PCMCI and run it to focus exclusively on contemporaneous (instantaneous) causal relationships. In the `Tigramite` library, this can be achieved by setting the maximum time lag τ_{max} to zero. This effectively disables the search for lagged causal dependencies. We select contemporary relationships as the ultimate result.

iv. **LPCMCI**: Similarly to PCMCI, we use partial correlation as the metric of CI test, and select the contemporary relationships as the obtained causal graph.

Study on Dimension of Latent Variables. We fix $d_x = 6$ and vary $d_z = \{2, 3, 4\}$ as shown in Table A6. The results indicate that both the Markov network and time-lagged structure are identifiable for lower dimensions. However, as the latent dimension increases, it witnesses a decline in the MCC, which is still the challenge in the continuous optimization of latent process identification [94, 43]. Nevertheless, the identifiability of latent space (R^2) remains satisfied across different settings.

Ablation Study on Conditions. We further conduct simulation studies to validate the theoretical identifiability guarantees under controlled settings with latent dimension $d_z = 3$ and observation dimension $d_x = 6$. To explicitly assess the necessity of key assumptions in our theory, we intentionally remove specific conditions, which are critical to the identifiability results. The following cases illustrate three distinct violations:

Table A6: **Results on Different Latent Dimensions.** We run simulations with 5 random seeds, selected based on the best-converged results to avoid local minima.

d_x	d_z	SHD ($\mathcal{G}_{x_t}$)	TPR	Precision	MCC ($\mathbf{s}_t$)	MCC ($\mathbf{z}_t$)	SHD ($\mathcal{G}_{z_t}$)	SHD ($\mathcal{M}_{lag}$)	R^2
6	2	0.12 ($\pm$0.04)	0.86 ($\pm$0.02)	0.85 ($\pm$0.04)	0.9864 ($\pm$0.01)	0.9741 ($\pm$0.03)	0.15 ($\pm$0.03)	0.21 ($\pm$0.05)	0.95 ($\pm$0.01)
	3	0.18 ($\pm$0.06)	0.83 ($\pm$0.02)	0.80 ($\pm$0.04)	0.9583 ($\pm$0.02)	0.9505 ($\pm$0.01)	0.24 ($\pm$0.06)	0.33 ($\pm$0.09)	0.92 ($\pm$0.01)
	4	0.23 ($\pm$0.02)	0.80 ($\pm$0.06)	0.74 ($\pm$0.01)	0.9041 ($\pm$0.02)	0.8931 ($\pm$0.03)	0.33 ($\pm$0.03)	0.48 ($\pm$0.05)	0.91 ($\pm$0.02)

Table A7: **Assumption Ablation Study.** These results verify the necessity of our assumptions in the theoretical analysis.

Setting	MCC ($\mathbf{s}_t$)	R^2
A	0.6328	0.34
B	0.7563	0.67
C	0.7052	0.85

i. **A** (Violation of contextual measurement condition in Theorem 1): We enforce conditional independence among $\mathbf{z}_t$ and replace the latent transition with an orthogonal mapping, thereby violating the 3-measurement condition required for block identifiability [24].

ii. **B** (Violation of Assumption 1): To violate the injectivity of linear operators, we use a simple autoregressive process $\mathbf{z}_t = \mathbf{z}_{t-1} + \boldsymbol{\epsilon}_{z_t}$ with $\boldsymbol{\epsilon}_{z_t} \sim \text{Uniform}(0, 1)$, which fails the injectivity requirement for $L_{z_t|z_{t-1}}$ and $L_{x_{t-1}|x_{t+1}}$ [53].

iii. **C** (Violation of Assumption 3): We constrain the generation variability by setting $\mathbf{s}_t = q(\mathbf{z}_t) + \boldsymbol{\epsilon}_{x_t}$, where q is a fixed mixing function and $\boldsymbol{\epsilon}_{x_t} \sim \mathcal{N}(0, \mathbf{I}_{d_x})$. This results in a linear Gaussian model without heteroscedasticity, undermining the necessary distributional variability, as discussed in [86].

As shown in Table A7, the removal of these assumptions leads to a substantial drop in both R^2 and MCC for $\mathbf{s}_t$, indicating a failure to recover the latent space and the observation-level causal structure. These findings empirically substantiate the necessity of our theoretical assumptions and delineate the conditions under which identifiability breaks down.

Hyperparameter Sensitivity. We test the hyperparameter sensitivity of CaDRe w.r.t. the sparsity and DAG penalty, as these hyperparameters have a significant influence on the performance of structure learning. In this experiment, we set $d_z = 3$ and $d_z = 6$. As shown in Table A8, the results demonstrate robustness across different settings, although the performance of structure learning is particularly sensitive to the sparsity constraint.

Table A8: **Hyperparameter Sensitivity.** We run experiments using 5 different random seeds for data generation and estimation procedures, reporting the average performance on evaluation metrics. "/" indicates loss of explosion. Notably, an excessively large DAG penalty at the beginning of training can result in a loss explosion or the failure of convergence.

α	1×10^{-5}	5×10^{-5}	1×10^{-4}	5×10^{-4}	1×10^{-3}	1×10^{-2}
SHD	0.23	0.22	0.18	0.27	0.32	0.67
β	1×10^{-5}	5×10^{-5}	1×10^{-4}	5×10^{-4}	1×10^{-3}	1×10^{-2}
SHD	0.37	0.18	0.20	/	/	/

E.3 On Real-world Dataset

Dataset Description.

1. Weather [2] dataset offers 10-minute summaries from an automated rooftop station at the Max Planck Institute for Biogeochemistry in Jena, Germany.

[2]https://www.bgc-jena.mpg.de/wetter/

Table A9: **Extended Results on Weather Forecasting.** Lower MSE/MAE is better. **Bold** numbers represent the best performance among the models, while underlined numbers denote the second-best.

Dataset	Length	CaDRe		iTransformer		Informer		PatchTST		DLinear+FAN		TimesNet		DLinear		N-Transformer	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	48	**0.125**	**0.167**	0.140	0.179	0.177	0.218	0.148	0.188	0.158	0.217	0.138	0.191	0.156	0.198	0.143	0.195
	96	**0.157**	**0.203**	0.168	0.214	0.225	0.259	0.187	0.226	0.199	0.256	0.180	0.231	0.186	0.229	0.199	0.246
	144	0.180	**0.225**	**0.172**	0.225	0.278	0.297	0.207	0.242	0.312	0.274	0.190	0.244	0.199	0.244	0.225	0.267
	192	0.207	**0.248**	**0.193**	0.241	0.354	0.348	0.234	0.265	0.238	0.298	0.212	0.265	0.217	0.261	2.960	0.315

2. CESM2 Pacific SST dataset employs monthly Sea Surface Temperature (SST) data generated from a 500-year pre-2020 control run of the CESM2 climate model. The dataset is restricted to oceanic regions, excluding all land areas, and retains its native gridded structure to preserve spatial correlations. It encompasses 6000 temporal steps, representing monthly SST values over the designated period. Spatially, the dataset comprises a grid with 186 latitude points and 151 longitude points, resulting in 28086 spatial variables, including 3337 land points where SST is undefined, and 24749 valid SST observations. To accommodate computational constraints, a downsampled version of the data, reduced to 84 grid points (6×14), is utilized in our experiment.
3. WeatherBench [62] is a benchmark dataset specifically tailored for data-driven weather forecasting. We specifically selected wind direction data for visualization comparisons within the same period, maintaining the original 350,640 timestamps.

Extended Weather Forecasting Results. To show the effectiveness of our approach, we evaluate additional large-scale time series forecasting models on the Weather dataset, a widely used benchmark in this domain. The selected models include iTransformer [48], PatchTST [57], Informer [46], DLinear [90], FAN [88], TimesNet [81], and N-Transformer [49]. As shown in Table A9, their results are reported to provide a comprehensive comparison with our method.

Initialization of Observational Causal Graph. To improve the stability of continuous optimization and avoid poor local minima in estimating the causal structure matrix $\hat{\mathcal{G}}_{\mathbf{x}_t}$, we incorporate a prior based on the Spatial Autoregressive (SAR) model. The SAR model, widely applied in geography, economics, and environmental science, captures spatial dependencies through the formulation:

$$\mathbf{X} = \mathbf{Z}\beta + \lambda \mathbf{W}\mathbf{X} + \mathbf{E},$$

where $\mathbf{W}$ is the spatial weights matrix, β is a regression coefficient, and $\mathbf{E}$ is a noise term. To simplify the model and isolate the spatial interaction component, we set $\beta = 0$ and $\lambda = 1$, resulting in the canonical SAR model:

$$\mathbf{X} = \mathbf{W}\mathbf{X} + \mathbf{E}.$$

The core assumption is that instantaneous interactions between regions are unlikely if they are separated by a substantial spatial distance. Therefore, we initialize $\mathbf{W}$ using a binary spatial adjacency matrix $\mathcal{M}_{\text{loc}}$, defined as

$$[\mathcal{M}_{\text{loc}}]_{i,j} = \mathbb{I}\left(\|s_i - s_j\|_2 \leq 50\right),$$

where s_i and s_j denote the spatial coordinates of regions i and j, respectively. This constraint enforces that only regions within a Euclidean distance of 50 units are considered spatially adjacent. We then estimate λ and update $\mathbf{W}$ by fitting the linear model $\mathbf{X} = \mathbf{W}\mathbf{X} + \mathbf{E}$ via least squares. The resulting matrix $\mathbf{W}$ is used as the initialization for the observational causal graph, denoted $\mathcal{M}_{\text{init}}$.

Compute the Observational Causal Graphs. Using a mask gradient-based approach, we compute an initial estimate of the Jacobian $\mathbf{J}_{\hat{g}}(\mathbf{x}_t)$, which encodes local sensitivities. However, these Jacobian matrices are typically dense and difficult to interpret directly. To produce a more interpretable visualization of the observational causal graph, we apply a masking operation followed by elementwise thresholding:

$$\hat{\mathcal{G}}_{x_t} = \mathbb{I}\left(|\mathbf{J}_{\hat{g}}(\mathbf{x}_t) \odot \mathcal{M}_{\text{init}}| > \tau\right), \tag{A53}$$

where $\odot$ denotes the elementwise (Hadamard) product, and $\mathbb{I}(\cdot)$ is the indicator function that outputs 1 if the condition is true and 0 otherwise. We set the threshold to $\tau = 0.15$ to obtain a binary adjacency matrix. $\mathcal{M}_{\text{init}}$ is the initialization mask.

To compute the partial Jacobian $\mathbf{J}_{\hat{g}}(\mathbf{x}_t)$ with respect to $\mathbf{s}_t$ while keeping $\mathbf{z}_t$ fixed, we disable gradient tracking for $\mathbf{z}_t$ by setting `requires_grad=False`, and use `autograd.functional.jacobian` in PyTorch.

Table A10: Architecture details. T, length of time series. $|\mathbf{x}_t|$: input dimension. n: latent dimension. LeakyReLU: Leaky Rectified Linear Unit. Tanh: Hyperbolic tangent function.

Configuration	Description	Output
ϕ	`z-encoder`	
Input:$\mathbf{x}_{1:t}$	Observed time series	Batch Size$\times T\times$ d_x
Dense	d_x neurons	Batch Size$\times T \times d_x$
Concat zero	concatenation	Batch Size$\times T \times d_x$
Dense	d_z neurons	Batch Size$\times T \times d_z$
η	`s-encoder`	
Input:$\mathbf{x}_{1:t}$	Observed time series	Batch Size$\times T\times$ d_x
Dense	d_x neurons	Batch Size$\times T \times d_x$
Concat zero	concatenation	Batch Size$\times T \times d_x$
Dense	d_x neurons	Batch Size$\times T \times d_x$
ψ	`decoder`	
Input:$\mathbf{z}_{1:T}$	Latent Variable	Batch Size$\times T \times (d_z + d_x)$
Dense	d_x neurons, Tanh	Batch Size$\times T \times d_x$
r	Modular Prior Networks	
Input: $\mathbf{z}_{1:T}$	Latent Variable	Batch Size$\times(d_z + 1)$
Dense	128 neurons,LeakyReLU	$(d_z + 1)\times$128
Dense	128 neurons,LeakyReLU	128$\times$128
Dense	128 neurons,LeakyReLU	128$\times$128
Dense	1 neuron	Batch Size$\times$1
Jacobian Compute	Compute log(det(J))	Batch Size

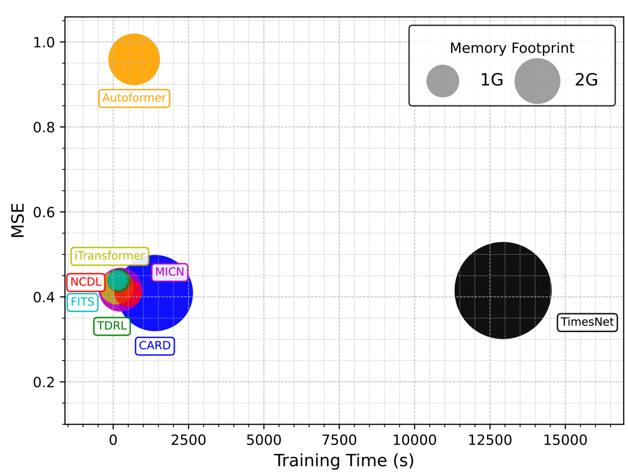


Figure A8: **Comparison of Computational Cost.** Different colors represent different methods, while the size of the circles corresponds to memory usage. The prediction length is set to 96.

Runtime and Computational Efficiency. We report the computational cost of the different methods considered. The comparison considers metrics including training time, memory usage, and corresponding performance MSE in the forecasting task. Note that inference time is not included in the comparison, as our work focuses on causal structure learning through continuous optimization rather than constraint-based methods. Figure A8 shows that our CaDRe method simultaneously learns the causal structure while achieving the lowest MSE, highlighting the importance of building a transparent and interpretable model. Furthermore, CaDRe exhibits similar training time and memory usage compared to mainstream time-series forecasting models in the lightweight track.

Model Structure We choose MICN [79] as the encoder backbone of our model on real-world datasets. Specifically, given that the MICN extracts the hidden feature, we apply a variational inference block and then an MLP-based decoder. Architecture details of the proposed method are shown in Table A10.

F More Discussions

F.1 Identifiability of Latent Space in n-order Markov Process

Theorem A2. ***(Identifiability of Latent Space in n-Order Markov Process)*** *Suppose observed variables and hidden variables follow the data-generating process in Eq.* (1)*, and estimated observations match the true joint distribution of* $\{\mathbf{x}_{t-n}, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_t, \ldots, \mathbf{x}_{t+n}, \mathbf{x}_{t+n+1}, \ldots, \mathbf{x}_{t+2n}\}$*. The following assumptions are imposed:*

- *A1' (Computable Probability:) The joint, marginal, and conditional distributions of* $(\mathbf{x}_t, \mathbf{z}_t)$ *are all bounded and continuous.*
- *A2' (Contextual Variability:) The operators* $L_{\mathbf{x}_{t+n+1:t+2n}|\mathbf{z}_{t:t+n}}$ *and* $L_{\mathbf{x}_{t-n:t-1}|\mathbf{x}_{t+n+1:t+2n}}$ *are injective and bounded.*
- *A3' (Latent Drift:) For any* $\mathbf{z}_{t:t+n}^{(1)}, \mathbf{z}_{t:t+n}^{(2)} \in \mathcal{Z}_t$ *where* $\mathbf{z}_{t:t+n}^{(1)} \neq \mathbf{z}_{t:t+n}^{(2)}$*, we have* $p(\mathbf{x}_t|\mathbf{z}_{t:t+n}^{(1)}) \neq p(\mathbf{x}_t|\mathbf{z}_{t:t+n}^{(2)})$*.*
- *A4' (Differentiability:) There exists a functional* M *such that* $M\left[p_{\mathbf{x}_{t:t+n}|\mathbf{z}_{t:t+n}}(\cdot \mid \mathbf{z}_{t:t+n})\right] = h_z(\mathbf{z}_{t:t+n})$ *for all* $\mathbf{z}_{t:t+n} \in \mathcal{Z}_{t:t+n}$*, where* h_z *is differentiable.*

Then we have $\hat{\mathbf{z}}_{t:t+n} = h_z(\mathbf{z}_{t:t+n})$*, where* $h_z : \mathbb{R}^{d_z \times n} \to \mathbb{R}^{d_z \times n}$ *is an invertible and differentiable function.*

If an n-order Markov process exhibits conditional independence across different time lags, block-wise identifiability of the conditioning variables can still be achieved using 3n measurements. For instance, when the lag is 2, once block-wise identifiability of the joint variables $[\mathbf{z}_t, \mathbf{z}_{t+1}]$ and $[\mathbf{z}_{t+1}, \mathbf{z}_{t+2}]$ is established, and given the known temporal direction $\mathbf{z}_t \to \mathbf{z}_{t+1}$, one can disambiguate $\mathbf{z}_t$ and $\mathbf{z}_{t+1}$ under mild variability assumptions. Subsequently, the same strategy as in Theorem A1 can be applied to achieve component-wise identifiability of $\mathbf{z}_t$ and $\mathbf{z}_{t+1}$ by leveraging conditional independencies given $\mathbf{z}_{t-2}$ and $\mathbf{z}_{t-1}$.

F.2 Allowing Time-Lagged Causal Relationships in Observations

In this section, we demonstrate that our proposed framework is compatible with the consideration of time-lagged effects, by providing *potential solutions*.

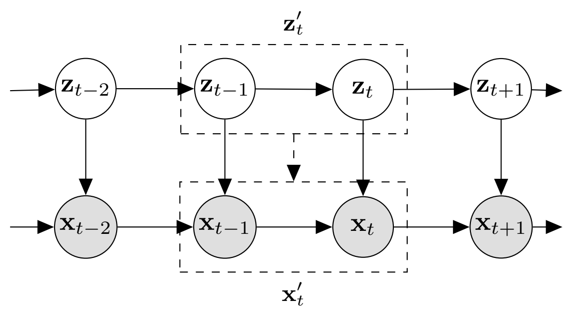


Figure A9: **4-measurement model with time-lagged effects in observed space.** $\mathbf{x}_t$ could be considered as the directed (dominating) measurement of $\mathbf{z}_t$, and $\mathbf{x}_{t-2}$, $\mathbf{x}_{t-1}$, and $\mathbf{x}_{t+1}$ provide indirect measurements of $\mathbf{z}_t$. For identifying the time-lagged causal relationships in observed space, we consider $\mathbf{z}'_t = (\mathbf{z}_{t-1}, \mathbf{z}_t)$ as the new latent variables, and $\mathbf{x}'_t = (\mathbf{x}_{t-1}, \mathbf{x}_t)$ as the new observed variables, to apply our *functional equivalence* in Theorem 2.

F.2.1 Phase I: Identifying Latent Variables from Time-Lagged Causally-Related Observations

For the identification of latent variables, we adopt the strategy outlined in [7, 25] to construct a spectral decomposition. We extend this approach to develop a proof strategy that establishes the identifiability of the latent space, as stated in Corollary A2.

We begin by defining the 4-measurement model, which includes time-series data with time-lagged effects in the observed space as a special case.

Definition 6 (4-Measurement Model)**.** $\mathbf{Z} = \{\mathbf{z}_{t-2}, \mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t+1}\}$ *represents latent variables in four continuous time steps, respectively. Similarly,* $\mathbf{X} = \{\mathbf{x}_{t-2}, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}\}$ *are observed variables that directly measure* $\mathbf{z}_{t-2}, \mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t+1}$ *using the same generating functions* g*. The model is defined by the following properties:*

- *The transformation within* $\mathbf{z}_{t-2}, \mathbf{z}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t+1}$ *is not measure-preserving.*
- *Joint density of* $\mathbf{x}_{t-2}, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{z}_t$ *is a product measure w.r.t. the Lebesgue measure on* $\mathcal{X}_{t-2} \times \mathcal{X}_{t-1} \times \mathcal{X}_t \times \mathcal{X}_{t+1} \times \mathcal{Z}_t$ *and a dominating measure* μ *is defined on* $\mathcal{Z}_t$*.*
- ***Limited feedback****:* $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{z}_t, \mathbf{z}_{t-1}) = p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{z}_t)$*.*
- *The distribution over* $(\mathbf{X}, \mathbf{Z})$ *is Markov and faithful to a DAG.*

Limited feedback explicitly assumes that future events do not cause past events and excludes instantaneous effects from $\mathbf{x}_t$ to $\mathbf{z}_t$. As illustrated in Figure A9, $\mathbf{x}_{t-2}, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}$ are defined as different measurements of $\mathbf{z}_t$, forming a temporal structure characteristic of a typical 4-measurement model. Under the data-generating process depicted in Figure A9, and based on the assumption of limited feedback, we propose the following framework:

$$\begin{aligned} p(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{x}_{t+2}) &= \int_{\mathcal{Z}_t} p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{z}_t) p(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{z}_t) p(\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \mathbf{z}_t) dz_t \\ &= \int_{\mathcal{Z}_t} p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, \mathbf{z}_t) p(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{z}_t) p(\mathbf{x}_{t-2} \mid \mathbf{z}_t, \mathbf{x}_{t-1}) dz_t. \end{aligned} \tag{A54}$$

Discussion of achieving the identifiability of latent space. Comparing Eq. A54 with Eq. A3, which represents the foundational result for proving the identifiability of latent space under the 3-measurement model, we extend the identification strategy from [7, 25] to the 4-measurement model. This forms the critical step in our identification process. We adopt assumptions analogous to those in [7, 25] and Theorem 1, and suppose the followings:

i. The joint distribution of $(\mathbf{X}, \mathbf{Z})$ and all their marginal and conditional densities are bounded and continuous.

ii. The linear operators $L_{x_{t+1}|x_t,z_t}$ and $L_{x_{t-2},x_{t-1},x_t,x_{t+1},z_t}$ are injective for bounded function space.

iii. For all $\mathbf{z}_t, \mathbf{z}'_t \in \mathcal{Z}_t$ $(\mathbf{z}_t \neq \mathbf{z}'_t)$, the set $\{\mathbf{x}_t : p(\mathbf{x}_t|\mathbf{z}_t) \neq p(\mathbf{x}_t|\mathbf{z}'_t)\}$ has positive probability.

hold true. Similar to the proof of our identifiability of latent space in Section A.2, except for the conditional independence introduced by the temporal structure, the key assumptions include an injective linear operator to enable the recovery of the density function of latent variables and distinctive eigenvalues to prevent eigenvalue degeneracy. The primary difference is the property *limited feedback*, where we can adopt the strategy in [7] to construct a unique spectral decomposition, where $(\mathbf{x}_{t-2}, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{z}_t)$ correspond to (X, S, Z, Y, X^*), respectively.
Following this, we apply the key steps of our identification process as detailed in the Appendix A.2. Ultimately, we can establish that the block $(\mathbf{z}_t, \mathbf{x}_t)$ is identifiable up to an invertible transformation:

$$(\hat{\mathbf{z}}_t, \hat{\mathbf{x}}_t) = h_{x,z}(\mathbf{z}_t, \mathbf{x}_t). \tag{A55}$$

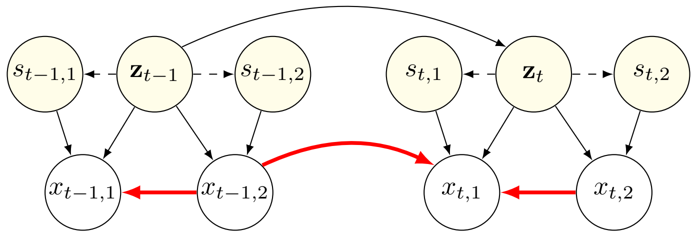
where $h_{x,z} : \mathbb{R}^{d_x+d_z} \to \mathbb{R}^{d_x+d_z}$ is a invertible function. Since the observation $\mathbf{x}_t$ is known and suppose $\hat{\mathbf{x}}_t = \mathbf{x}_t$, this relationship indeed represents an invertible transformation between $\hat{\mathbf{z}}_t$ and $\mathbf{z}_t$ as

$$\hat{\mathbf{z}}_t = h_z(\mathbf{z}_t). \tag{A56}$$

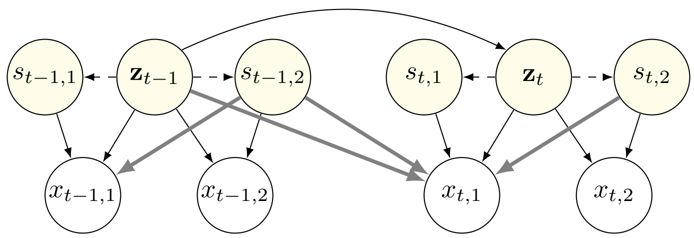
With an additional assumption of a sparse latent Markov network, we achieve component-wise identifiability of the latent variables, as stated in Theorem A1 in appendix, leveraging the proof strategies of [94, 43]. These results are stronger than those in [7].

F.2.2 Phase II: Identifying Time-Lagged Observation Causal Graph

Unified Modeling across Neighboring Time Points. In the presence of time-lagged effects in the observed space, such as $\mathbf{x}_{t-1} \to \mathbf{x}_t$, alongside the causal DAG within $\mathbf{x}_t$, as depicted in Figure A9, we show that by introducing an expanded set of latent variables $\mathbf{z}'_t = (\mathbf{z}_{t-1}, \mathbf{z}_t)$ and an expanded set of observed variables $\mathbf{x}'_t = (\mathbf{x}_{t-1}, \mathbf{x}_t)$, the property of functional equivalence is preserved. Moreover, identifiability continues to hold, and, broadly speaking, it becomes more accessible due to the incorporation of Granger causality principles in time-series data [16], if we assume that future events cannot influence or cause past events.



(a) Time-lagged SEM.



(b) Equivalent ICA.

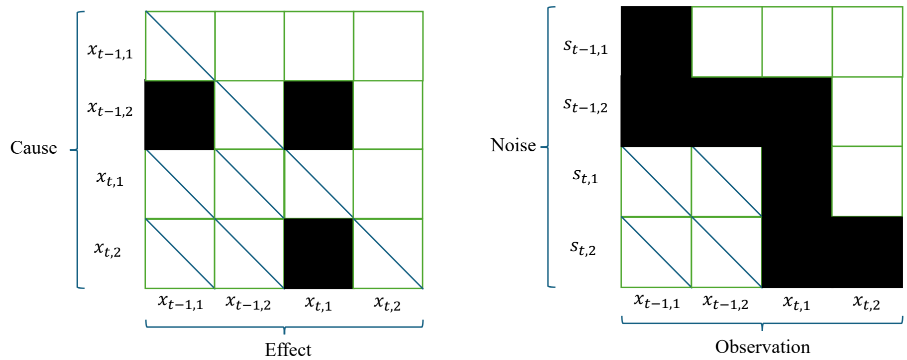
Figure A10: Equivalent time-lagged SEM and ICA in the case with time-lagged causal relationships in observed space. The red lines in Figure A10a indicate that information are transmitted by the instantaneous and the time-lagged observational causal graphs, while the gray lines in Figure A10b represent that the information transitions are equivalent to originating from contemporary $\mathbf{s}_t$ and previous $(\mathbf{z}_t, s_{t-1,2})$ within the mixing structure.

Functional Equivalence in Presence of Time-Lagged Effects. As shown in Figure A10, we show that, if we consider the time-lagged causal relationship in observed space, it still can be processed with the technique as in our paper proposed, through considering time-lagged causal relationships as a part of observational causal graph, by reformulating $\mathbf{z}'_t = (\mathbf{z}_{t-1}, \mathbf{z}_t)$, $\mathbf{x}'_t = (\mathbf{x}_{t-1}, \mathbf{x}_t)$ and $\mathbf{s}'_t = (\mathbf{s}_{t-1}, \mathbf{s}_t)$, to apply the Corollary A2. Specifically, the time-lagged effects from $\mathbf{x}_{t-2}$ can be considered as side information, which does not make difference to causal relationships from $\mathbf{x}_{t-1}$ to $\mathbf{x}_t$ and its corresponding ICA form.

F.2.3 Estimation Methodology

Slided Window. Building on the analysis above, we aggregate two adjacent time-indexed observations into a single new observation. By employing a sliding window with a step size of 1, we obtain $T-1$ new observations along with their corresponding latent variables, thereby aligning with the estimation methodology described in Section C.

Structure Prunning. For structure learning, given the assumption that future climate cannot cause past climate, we can mask $\frac{1}{4}$ of elements in the causal adjacency matrix during implementation, as depicted in Figure A11. Compared with the original implementation, the masking simplifies the difficulty of optimization by reducing the degrees of freedom in the graph.



(a) Causal adjacency matrix of SEM. (b) Mixing matrix of equivalent ICA.

Figure A11: Interpreting Figure A10 with causal adjacency matrix of the SEM and the mixing matrix of the equivalent ICA. The diagonal lines indicate masked elements, as future events cannot cause past events, and self-loops are not permitted. Black blocks represent the presence of a causal relationship or functional dependency in the generating function g_m, while white blocks indicate the absence of such a relationship.

G Broader Impacts

The proposed CaDRe framework offers a substantial advancement in climate science by enabling the joint identification of latent dynamic processes and observational causal structures from purely observational data. Understanding these structures is critical for interpreting complex atmospheric phenomena, improving forecasting accuracy, and informing climate-related decision-making. By providing identifiability guarantees without relying on restrictive assumptions, CaDRe addresses fundamental limitations in existing climate modeling approaches, particularly in the presence of latent confounders and observational noise.

The ability to recover interpretable latent drivers and causal graphs directly from climate data enhances scientific understanding and supports more transparent and robust climate models. This is especially valuable for anticipating and responding to climate variability and extreme events. Moreover, the theoretical framework underlying CaDRe extends to other scientific domains involving spatiotemporal processes, but its primary impact lies in improving the causal interpretability and empirical grounding of climate analyses. As such, CaDRe represents a step toward causally principled climate modeling, with the potential to inform both scientific inquiry and policy development.

References

[1] Kashif Abbass, Muhammad Zeeshan Qasim, Huaming Song, Muntasir Murshed, Haider Mahmood, and Ijaz Younis. A review of the global climate change impacts, adaptation, and sustainable mitigation measures. *Environmental Science and Pollution Research*, 29(28):42539–42559, 2022.

[2] Lazar Atanackovic, Alexander Tong, Bo Wang, Leo J Lee, Yoshua Bengio, and Jason S Hartford. Dyngfn: Towards bayesian inference of gene regulatory networks with gflownets. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. In *International Conference on Machine Learning*, pages 899–908. PMLR, 2018.

[4] Tom Beucler and et al. Climatenet: Bringing the power of deep learning to climate science at scale. *arXiv preprint arXiv:2101.07148*, 2021.

[5] Julien Boé and Laurent Terray. Land–sea contrast, soil-atmosphere and cloud-temperature interactions: interplays and roles in future summer european climate change. *Climate dynamics*, 42(3):683–699, 2014.

[6] Philippe Brouillard, Sébastien Lachapelle, Julia Kaltenborn, Yaniv Gurwicz, Dhanya Sridhar, Alexandre Drouin, Peer Nowack, Jakob Runge, and David Rolnick. Causal representation learning in temporal data via single-parent decoding. *arXiv preprint arXiv:2410.07013*, 2024.

[7] Raymond J Carroll, Xiaohong Chen, and Yingyao Hu. Identification and estimation of nonlinear models using two samples with nonclassical measurement errors. *Journal of nonparametric statistics*, 22(4):379–399, 2010.

[8] Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, and Kun Zhang. Caring: Learning temporal causal representation under non-invertible generation process. *arXiv preprint arXiv:2401.14535*, 2024.

[9] Yi-Leng Chen and Jian-Jian Wang. The effects of precipitation on the surface temperature and airflow over the island of hawaii. *Monthly weather review*, 123(3):681–694, 1995.

[10] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[11] John B Conway. *A course in functional analysis*, volume 96. Springer Science & Business Media, 1994.

[12] Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. *arXiv preprint arXiv:2312.11001*, 2023.

[13] Nelson Dunford and Jacob T. Schwartz. *Linear Operators*. John Wiley & Sons, New York, 1971.

[14] Imme Ebert-Uphoff and Yi Deng. Causal discovery for climate research using graphical models. *Journal of Climate*, 25(17):5648–5665, 2012.

[15] Franklin M Fisher. A correspondence principle for simultaneous equation models. *Econometrica: Journal of the Econometric Society*, pages 73–92, 1970.

[16] John R Freeman. Granger causality and the times series analysis of political relationships. *American Journal of Political Science*, pages 327–358, 1983.

[17] Andreas Gerhardus and Jakob Runge. High-recall causal discovery for autocorrelated time series with latent confounders. *Advances in Neural Information Processing Systems*, 33:12615–12625, 2020.

[18] Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in neural information processing systems*, 34:28233–28248, 2021.

[19] Shiyu Gu, Tim Januschowski, and Jan Gasthaus. Efficiently modeling time series with missing data using a state space approach. In *NeurIPS Time Series Workshop*, 2021.

[20] Shiyu Gu, David Salinas, Valentin Flunkert, and Jan Gasthaus. Combining latent state-space models and structural time series models for probabilistic forecasting. *International Journal of Forecasting*, 37(3):1182–1199, 2021.

[21] Shiyu Gu, David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Parameterization of state space models for forecasting with structured latent dynamics. *arXiv preprint arXiv:2202.09384*, 2022.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[23] 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.

[24] Yingyao Hu and Susanne M Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

[25] Yingyao Hu and Matthew Shum. Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics*, 171(1):32–44, 2012.

[26] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

[27] Aapo Hyvärinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10), 2023.

[28] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

[29] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

[30] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[31] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[32] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[33] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[34] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

[35] Oleksandr Klushyn and et al. Latent-space forecasting of climate variables using variational autoencoders. *arXiv preprint arXiv:2107.01227*, 2021.

[36] Lingjing Kong, Biwei Huang, Feng Xie, Eric Xing, Yuejie Chi, and Kun Zhang. Identification of nonlinear latent hierarchical models. *Advances in Neural Information Processing Systems*, 36:2010–2032, 2023.

[37] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.

[38] Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024.

[39] Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

[40] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 95–104, 2018.

[41] Remi Lam and et al. Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*, 2022.

[42] Jan Lemeire and Dominik Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, 23:227–249, 2013.

[43] Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the identification of temporal causal representation with instantaneous dependence. In *The Thirteenth International Conference on Learning Representations*, 2025.

[44] Juan Lin. Factorizing multivariate function classes. *Advances in neural information processing systems*, 10, 1997.

[45] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. *arXiv preprint arXiv:2206.06169*, 2022.

[46] Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-tao Xia. Timebridge: Non-stationarity matters for long-term time series forecasting. *arXiv preprint arXiv:2410.04442*, 2024.

[47] 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.

[48] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

[49] Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35:9881–9893, 2022.

[50] Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. *arXiv preprint arXiv:2307.02390*, 2023.

[51] Valerio Lucarini, Richard Blender, Corentin Herbert, Francesco Ragone, Salvatore Pascale, and Jeroen Wouters. Mathematical and physical ideas for climate science. *Reviews of Geophysics*, 52(4):809–859, 2014.

[52] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

[53] Lutz Mattner. Some incomplete but boundedly complete location families. *The Annals of Statistics*, pages 2158–2162, 1993.

[54] Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in artificial intelligence*, pages 186–195. PMLR, 2020.

[55] Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. *arXiv preprint arXiv:2310.15709*, 2023.

[56] Ignavier Ng, Shengyu Zhu, Zhuangyan Fang, Haoyang Li, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 424–432. SIAM, 2022.

[57] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[58] Jaideep Pathak and et al. Fourcastnet: Global medium-range weather forecasting with graph neural networks. *arXiv preprint arXiv:2202.11214*, 2022.

[59] Judea Pearl. *Causality*. Cambridge university press, 2009.

[60] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[61] Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models. *arXiv preprint arXiv:2402.09236*, 2024.

[62] Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

[63] Markus Reichstein and et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019.

[64] Patrik Reizinger, Yash Sharma, Matthias Bethge, Bernhard Schölkopf, Ferenc Huszár, and Wieland Brendel. Jacobian-based causal discovery with nonlinear ica. *Transactions on Machine Learning Research*, 2023.

[65] Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

[66] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022.

[67] Jakob Runge. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, pages 1388–1397. Pmlr, 2020.

[68] Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

[69] 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.

[70] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. In *International Journal of Forecasting*, volume 36, pages 1181–1191. Elsevier, 2020.

[71] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[72] 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.

[73] Alessio Spantini, Daniele Bigoni, and Youssef Marzouk. Inference via low-dimensional couplings. *The Journal of Machine Learning Research*, 19(1):2639–2709, 2018.

[74] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[75] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2001.

[76] Adolf Stips, Diego Macias, Clare Coughlan, Elisa Garcia-Gorriz, and X San Liang. On the causal structure between co2 and global temperature. *Scientific reports*, 6(1):21691, 2016.

[77] Benjamin A. Toms and Elizabeth A. Barnes. Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, 12(12), 2020.

[78] Robert Vautard, Geert Jan Van Oldenborgh, Friederike EL Otto, Pascal Yiou, Hylke De Vries, Erik Van Meijgaard, Andrew Stepek, Jean-Michel Soubeyroux, Sjoukje Philip, Sarah F Kew, et al. Human influence on european winter wind storms such as those of january 2018. *Earth System Dynamics*, 10(2):271–286, 2019.

[79] Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. Micn: Multi-scale local and global context modeling for long-term series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022.

[80] Xue Wang, Tian Zhou, Qingsong Wen, Jinyang Gao, Bolin Ding, and Rong Jin. Card: Channel aligned robust blend transformer for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2023.

[81] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

[82] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

[83] Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with \$10k\$ parameters. In *The Twelfth International Conference on Learning Representations*, 2024.

[84] Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying causal representation learning with dynamical systems for science. *arXiv preprint arXiv:2405.13888*, 2024.

[85] Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.

[86] Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022.

[87] Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

[88] Weiwei Ye, Songgaojun Deng, Qiaosha Zou, and Ning Gui. Frequency adaptive normalization for non-stationary time series forecasting. *arXiv preprint arXiv:2409.20371*, 2024.

[89] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

[90] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

[91] Jiji Zhang. A comparison of three occam's razors for markovian causal models. *The British journal for the philosophy of science*, 2013.

[92] Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

[93] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

[94] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

[95] 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.

[96] Yujia Zheng, Ignavier Ng, Yewen Fan, and Kun Zhang. Generalized precision matrix for scalable estimation of nonparametric markov networks. *arXiv preprint arXiv:2305.11379*, 2023.

[97] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022.

[98] Yujia Zheng and Kun Zhang. Generalizing nonlinear ica beyond structural sparsity. *Advances in Neural Information Processing Systems*, 36:13326–13355, 2023.

[99] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.