Learning General Causal Structures with Hidden Dynamic Process for Climate Analysis

Minghao Fu

Ph.D. Student at UCSD

Background: Infer Causality in Climate System

- Understanding the causal structure of climate systems is fundamental not only to scientific reasoning, but also to reliable modeling and prediction.

Weather forecasting

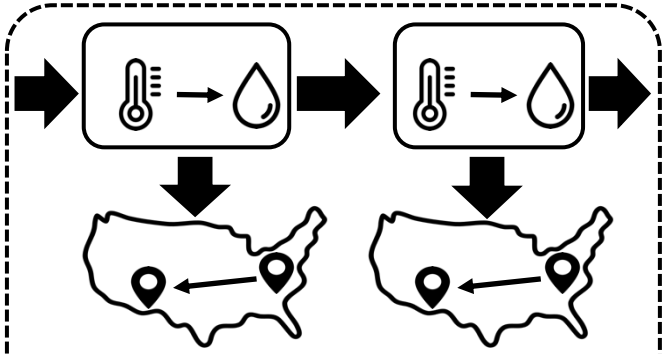


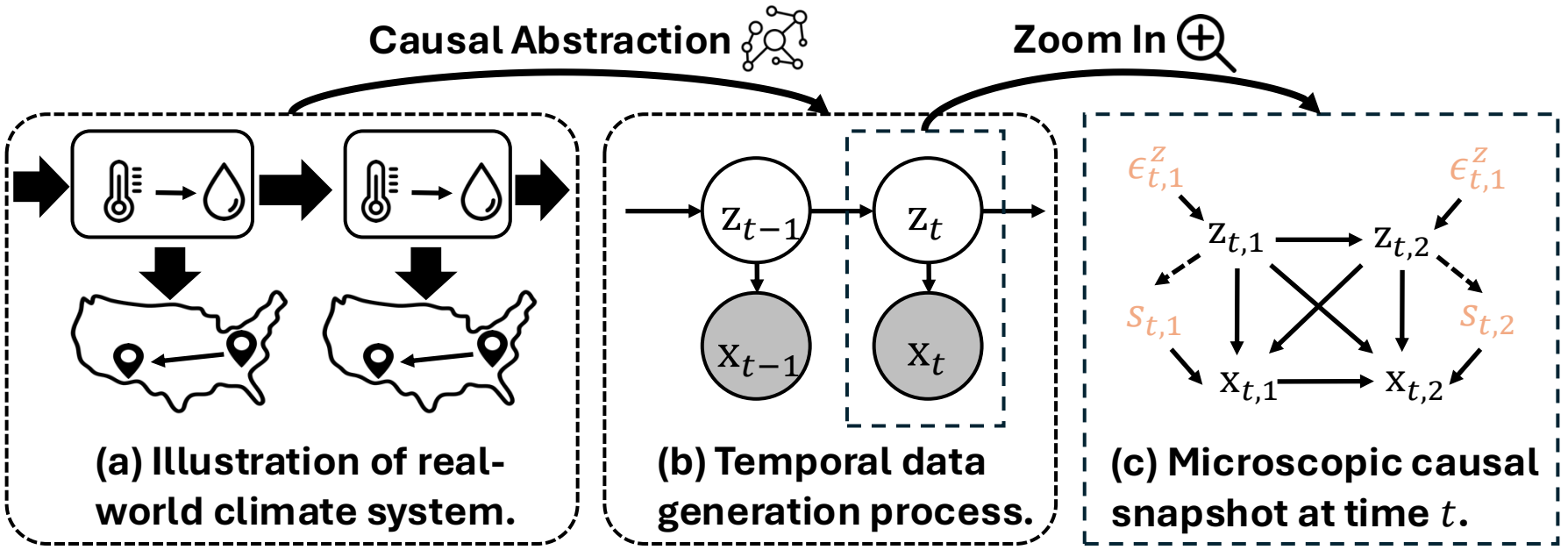
Illustration of real-world climate system.

- By uncovering how various components of the climate interact, such as atmospheric dynamics, ocean circulation, and biospheric feedbacks, researchers can distinguish causal relations from observable correlations.

Background: Infer Causality in Climate System

- Some climate variables are not directly measured but significantly influence the observed dynamics, exhibits both instantaneous and time-lagged causal dependencies.

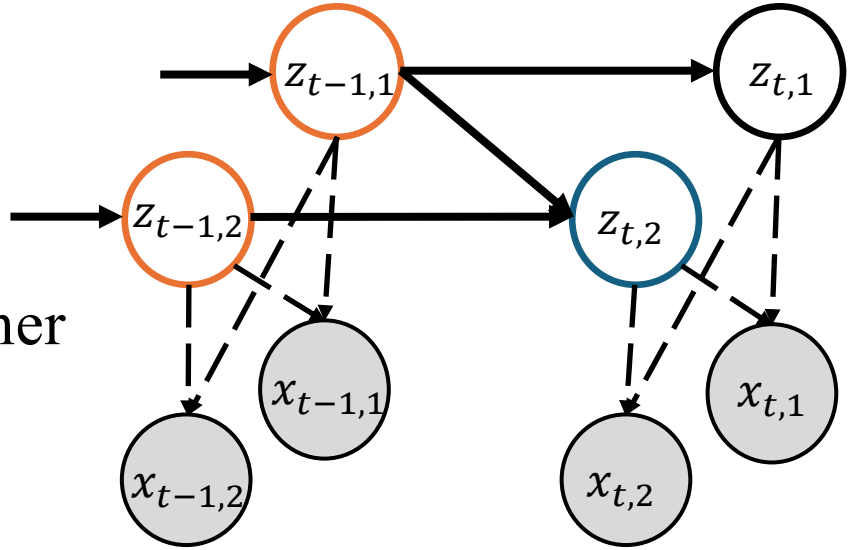
- Exhibit spatial interactions through emergent weather patterns, like wind circulation systems.



A Fundamental Problem in Causal Discovery:

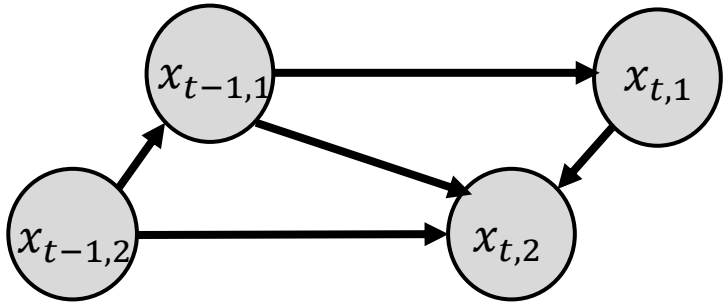
How to Jointly Recover Causal Factors in Hidden & Observed Space?

Prior methods focus on either



Latent Causal Process

or

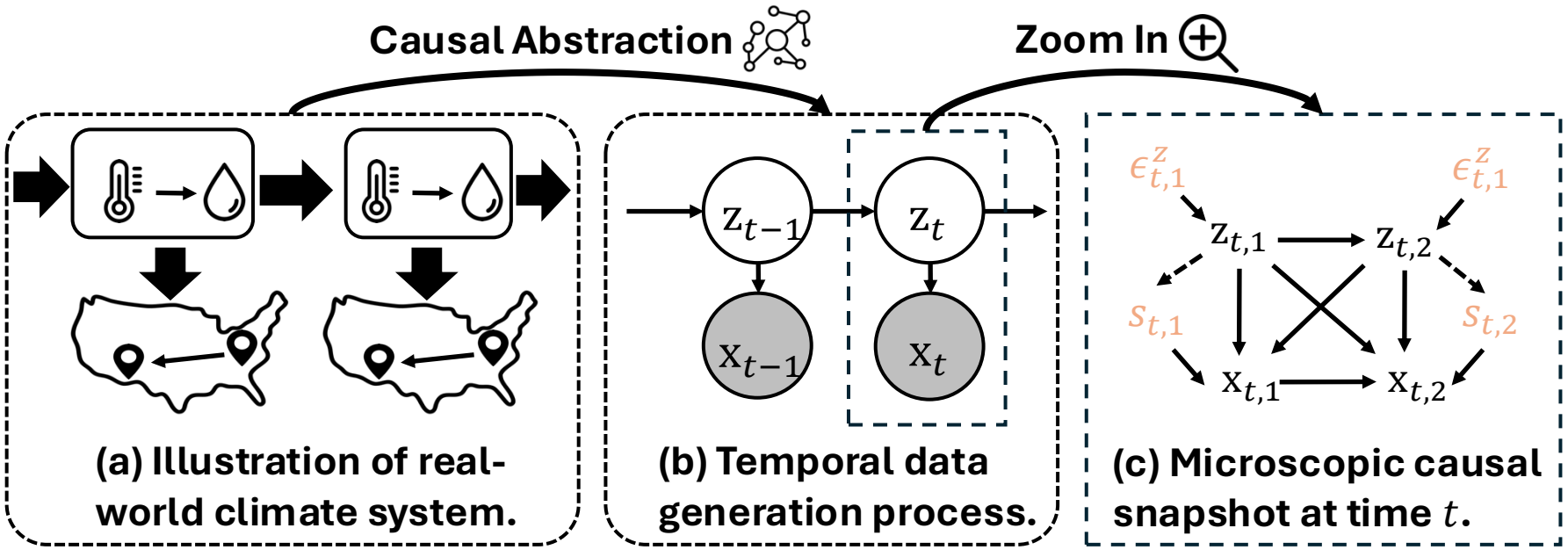


Causal Relations in Observed Space

➢ ***In a realistic climate system, both are present.***

A Unified Framework for Unveiling the Climate System

- In the temporal scenario, we introduce a **unified framework** CaDre that jointly uncover
  - (i) Causal relations among observed variables
  - (ii) Latent driving forces together with their interactions



Our Solution

Theoretical Foundation

What information and condition can unveil latent factors and causal relations?

Implementation

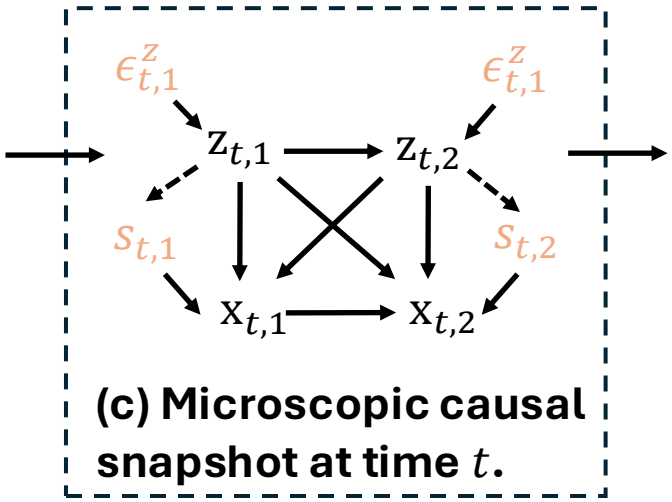
How can these theoretical foundations guide the design of model and objective function?

Problem Formulation

- We translate how the dynamic climate system evolves to the following Structural Equation Model (SEM) at each discrete time step:

$$x_{t,i} = \underbrace{g_i(\mathbf{pa}_O(x_{t,i}), \mathbf{pa}_L(x_{t,i}), s_{t,i})}_{\text{effects from } \mathbf{x}_t \text{ and } \mathbf{z}_t}, \quad z_{t,j} = \underbrace{f_j(\mathbf{pa}_L(z_{t,j}), \epsilon^z_{t,j})}_{\text{effects from } \mathbf{z}_{t-1} \text{ and } \mathbf{z}_t}, \quad \underbrace{s_{t,i} = g_{s_i}(\mathbf{z}_t, \epsilon^x_{t,i})}_{\text{noise conditioned on } \mathbf{z}_t},$$

- Causal relationships exist both in **observed space & latent process**
- The causal graphs can change **dynamically** with the hidden process
- All causal relations are **nonparametric**



(c) Microscopic causal snapshot at time t.

Emerged Problems in Identification

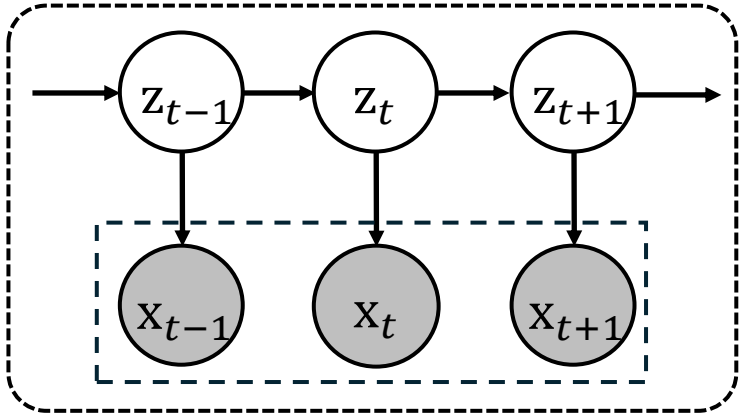
- Based on this formulation, two fundamental identification problem emerged:
  - (i) How to recover the latent space when observed variables are **causally-related**?
  - (ii) How to identify **nonparametric causal graphs** over observed variables, even if they are modulated by a hidden dynamic process?

$$x_{t,i} = \underbrace{g_i(\mathbf{pa}_O(x_{t,i}), \mathbf{pa}_L(x_{t,i}), s_{t,i})}_{\text{effects from } \mathbf{x}_t \text{ and } \mathbf{z}_t}, \quad z_{t,j} = \underbrace{f_j(\mathbf{pa}_L(z_{t,j}), \epsilon^z_{t,j})}_{\text{effects from } \mathbf{z}_{t-1} \text{ and } \mathbf{z}_t}, \quad \underbrace{s_{t,i} = g_{s_i}(\mathbf{z}_t, \epsilon^x_{t,i})}_{\text{noise conditioned on } \mathbf{z}_t},$$

- To formalize the stochastic generation process, we introduce an operator L (Dunford & Schwartz, 1971) to represent distribution-level transformations:

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

Recover Latent Space



- We extend the **3-measurement theory** (Hu & Schennach 2008) to temporal causal representation learning with causally-related observations:

Theorem 1. ***(Identifiability of Latent Space)*** *Suppose observed variables and hidden variables follow the data generating process in Eq. (1), and estimated observations* $\{\hat{\mathbf{x}}_{t-1}, \hat{\mathbf{x}}_t, \hat{\mathbf{x}}_{t+1}\}$ *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.*

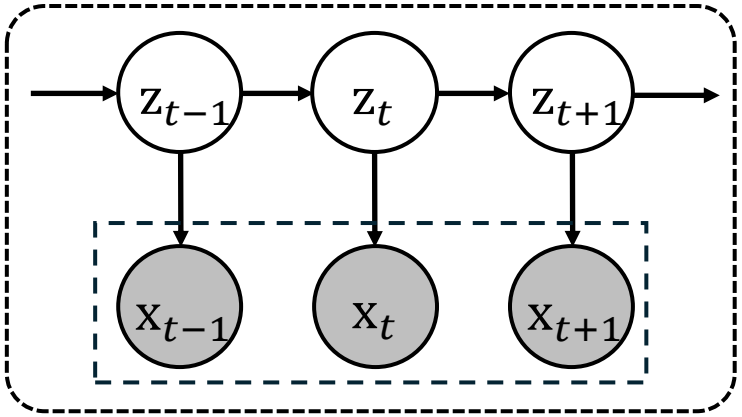
A3 (Latent Drift:) For any $\mathbf{z}_t^{(1)}, \mathbf{z}_t^{(2)} \in \mathcal{Z}_t$ *where* $\mathbf{z}_t^{(1)} \neq \mathbf{z}_t^{(2)}$, *we have* $p(\mathbf{x}_t|\mathbf{z}_t^{(1)}) \neq p(\mathbf{x}_t|\mathbf{z}_t^{(2)})$.

A4 (Differentiability:) There exists a functional F *such that* $F\left[p_{\mathbf{x}_t|\mathbf{z}_t}(\cdot \mid \mathbf{z}_t)\right] = h_z(\mathbf{z}_t)$ *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.*

Recover Latent Space: Proof Sketch and Contributions

- Contribution: Result of Hu & Schennach relies on partially knowing the function form of g, and yields only distribution-level identifiability.



- In contrast, our approach requires no such prior knowledge and achieves identifiability at the value level, an informative result for component-wise analysis.

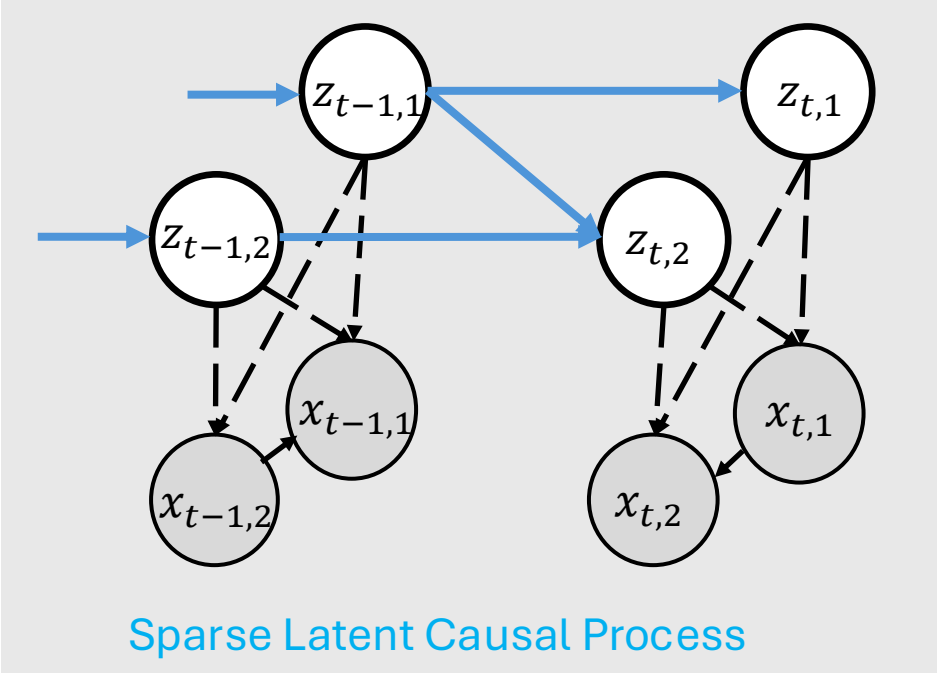
- **Proof Sketch:**

$$\{p_{\mathbf{x}_t|\mathbf{z}_t}(\cdot \mid \mathbf{z}_t)\}_{\mathcal{Z}_t} = \{p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\cdot \mid \hat{\mathbf{z}}_t)\}_{\hat{\mathcal{Z}}_t} \Rightarrow p_{\mathbf{x}_t|\mathbf{z}_t}(\mathbf{x}_t \mid h_z(\mathbf{z}_t)) = p_{\mathbf{x}_t|\hat{\mathbf{z}}_t}(\mathbf{x}_t \mid \hat{\mathbf{z}}_t) \Rightarrow \hat{\mathbf{z}}_t = h_z(\mathbf{z}_t)$$

(Hu & Schennach 2008) **(Theorem 1)**

Extend to Component-wise Identifiability

- For a scientific analysis, we introduce a sparsity assumption (Li et al., 2024) on the latent dynamics to achieve component-wise identifiability of latent variables and causal process.



Sparse Latent Causal Process

- Motivated by that physical climate factors, e.g., solar radiation and atmospheric, exhibit localized sparse influences.

DAG Invert Linear Operator

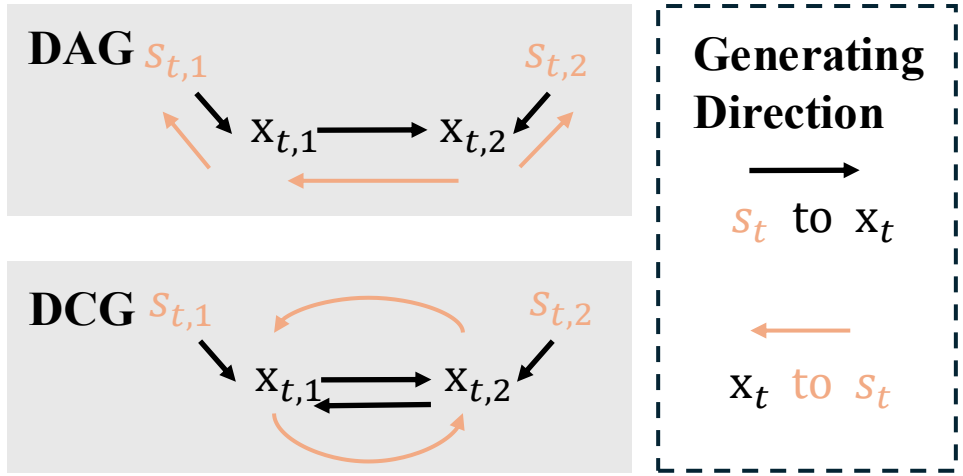
- Aiming at reliably discovering causal graphs, we adopt a widely-used assumption in causal discovery:

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

- If the causal graph is a DAG, causal influence can be traced back to its source by following the reverse direction of the DAG, instantiated as an injective linear operator:

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

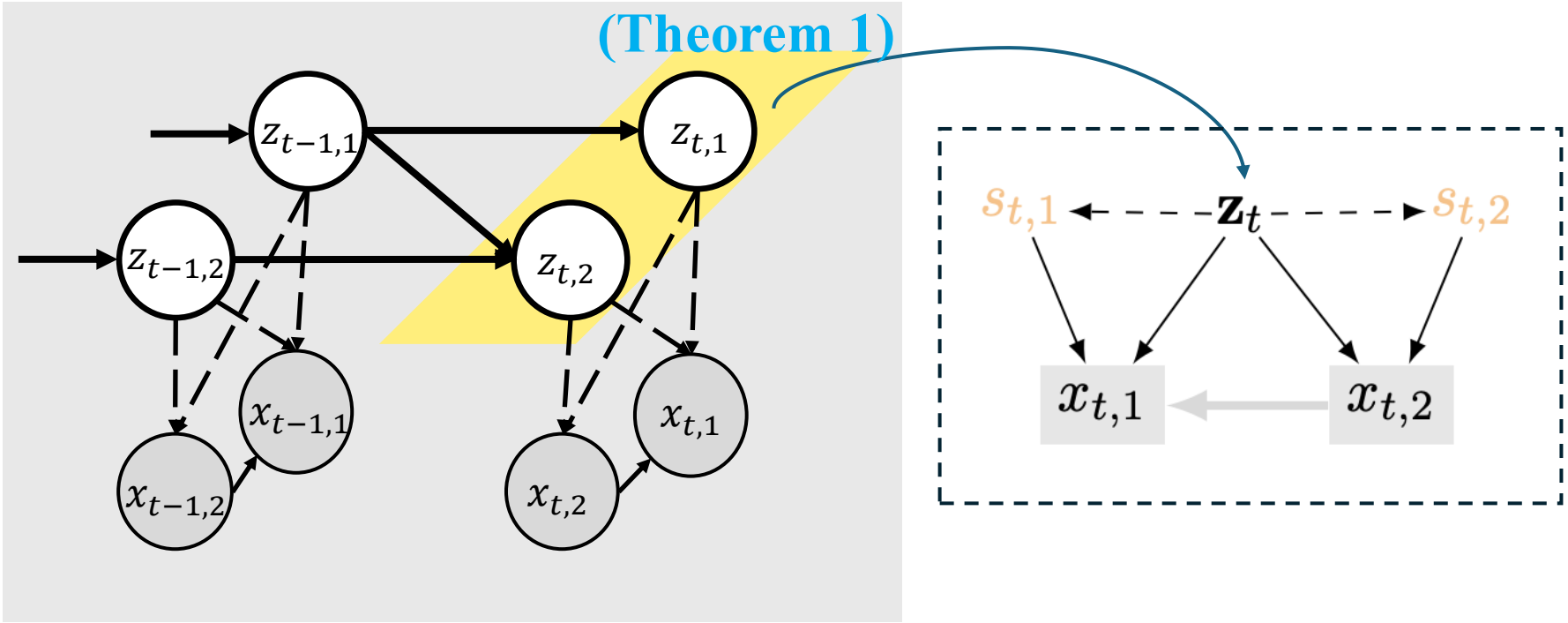
- ***DAG among observed variables does not disturb the injectivity of operator!***



Nonparametric Causal Discovery with Hidden Process

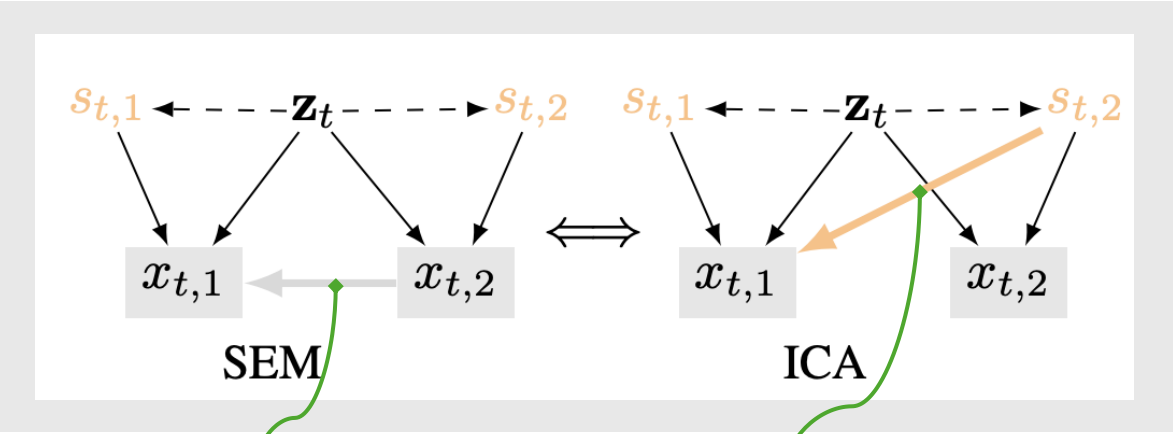
- Nonparametric causal graph with latents are not identifiable without auxiliary variables.

- Insight: The identified **hidden process** in prior results, can serve as auxiliary information!



Connecting SEM with Latents to nonlinear ICA

- Nonparametric causal discovery with latent variables is ill-posed.
- With auxiliary variables, a generative model (ICA) is more tractable for both identifiability analysis and practical implementations.



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

Eq. **Functional Equivalence**

- Identify the ICA model to equivalently **infer the causal graphs**.

Identifiability of Nonparametric Dynamic Causal Graphs

- In presence of auxiliary variables, nonlinear ICA are generally identifiable under sufficient variability. Faithfulness ensures the structural translation from ICA to SEM.

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

- Proof Sketch:

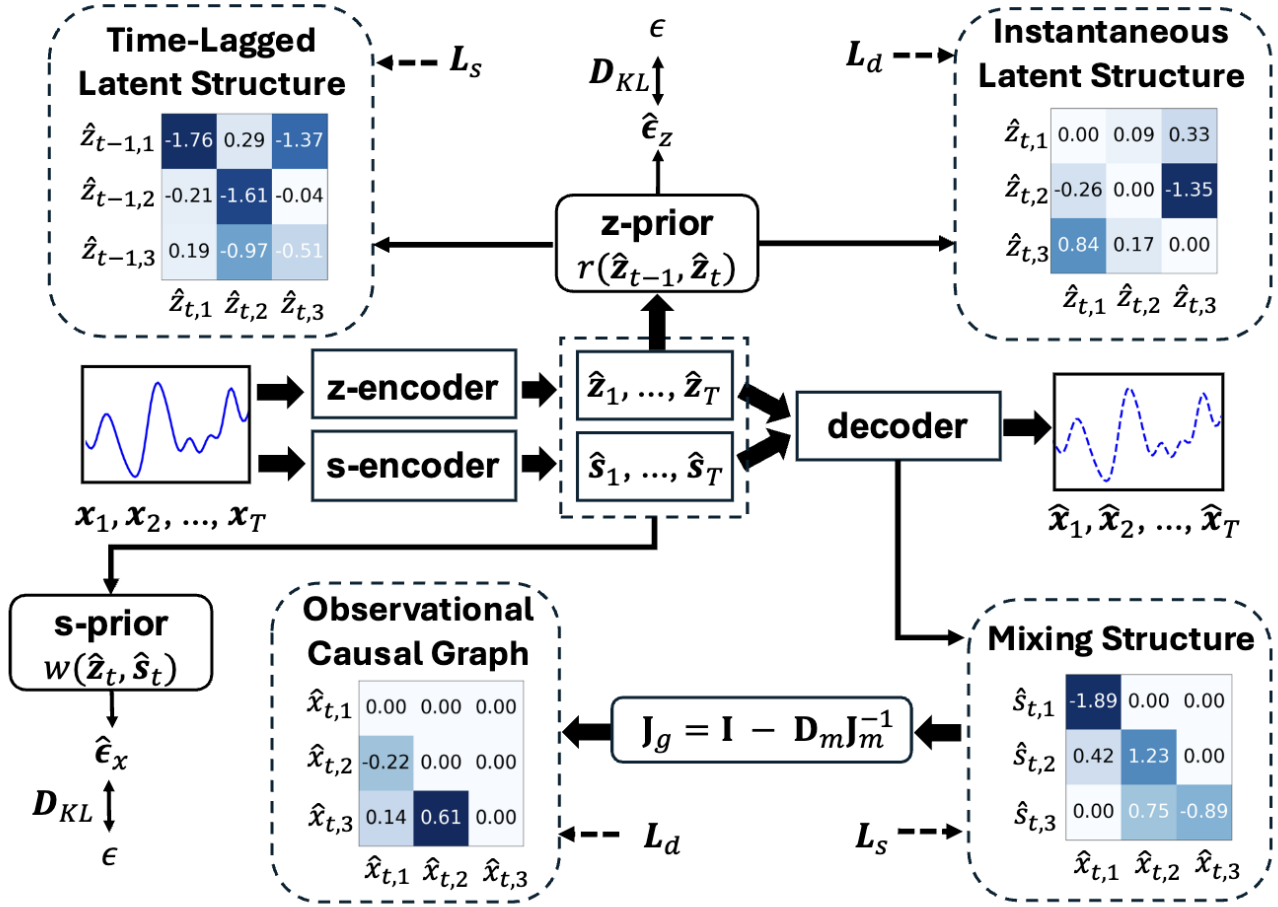
$$\boxed{\hat{\mathbf{z}}_t = h_z(\mathbf{z}_t)} \Rightarrow \boxed{\hat{s}_{t,i} = h_s(s_{t,\pi(i)})} \Rightarrow \boxed{\text{supp}(\mathbf{J}_{\hat{m}}) = \text{supp}(\mathbf{J}_m)} \Rightarrow \boxed{\text{supp}(\mathbf{J}_{\hat{g}}) = \text{supp}(\mathbf{J}_g)}$$

(Theorem 1) **(Nonlinear ICA)** **(DAG Constraint)** **(Infer Causal Graphs)**

- To eliminate this permutation ambiguity, we further exploit the structural constraints encoded by the DAG over observed variables.

Methodology

- z-encoder for extracting latent variables z_t, and s-encoder for extracting nonstationary noise s_t.
- A decoder reconstructs x_t from them.
- Prior networks estimate the prior distribution to learn the causal structures based on the **Jacobian matrix**.
- D_{KL} enforces independence of the estimated noise by minimizing its KL divergence w.r.t. $N(0, \mathbf{I})$



Overall Architecture

- ELBO: z-encoder for extracting latent variables z_t, and s-encoder for extracting nonstationary noise s_t. A decoder reconstructs x_t from them.

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

- z-encoder, s-encoder and decoder are implemented by Multi-Layer Perceptrons (MLPs) as:

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

Estimating z_t with s_t using Flow Networks.

- Prior networks estimate the prior distribution to learn the causal structures based on the **Jacobian matrix**.

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

- Using the same estimation backbone, we minimize the KL divergence between the prior of s_t and the posterior obtained from the s-encoder

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

Structure Learning

- To prevent redundant edges and cycles, a sparsity penalty is imposed on each structure:

$$\mathcal{L}_s = \|\mathcal{M}(\mathbf{J}_r(\hat{\mathbf{z}}_t))\|_1 + \|\mathcal{M}(\mathbf{J}_d(\hat{\mathbf{z}}_{t-1}))\|_1 + \|\mathbf{J}_{\hat{g}}(\hat{\mathbf{x}}_t)\|_1$$

- DAG constraints are imposed on the causal graph over observed variables and instantaneous latent causal DAG

$\mathcal{L}_d = \mathcal{D}_g(\mathbf{J}_{\hat{g}}(\hat{\mathbf{x}}_t)) + \mathcal{D}_g(\mathbf{J}_r(\hat{\mathbf{z}}_t))$, where $\mathcal{D}_g(A) = \operatorname{tr}\left[(I + \frac{1}{d}A \circ A)^d\right] - d$

Experiment Results: Comparison with Constraint-based CD

➢ CaDRe outperforms all baselines across different sample sizes and variable dimensions.

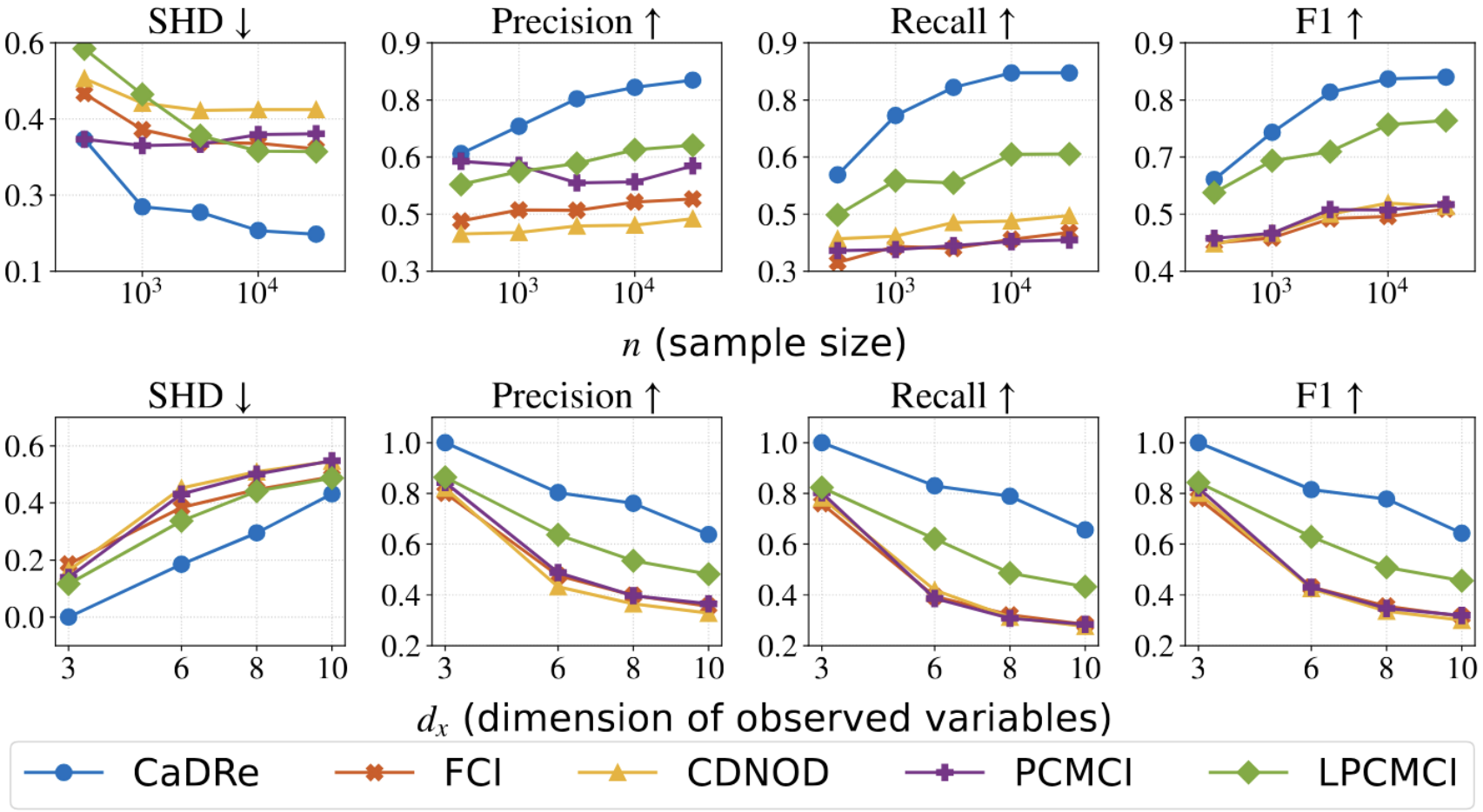


Fig. Causal discovery results compared with constraint-based method

Experiment Results: Comparison with Temporal CRL

- Considered methods fail to recover latent variables, as they cannot properly address cases where the observed variables are causally-related.

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

Tab. Identification results compared with temporal CRL

Experiment Results: Climate Forecasting

- Our approach outperforms existing time-series forecasting models in precision, due to existing models struggling with causally-related observations and non-contaminated generation

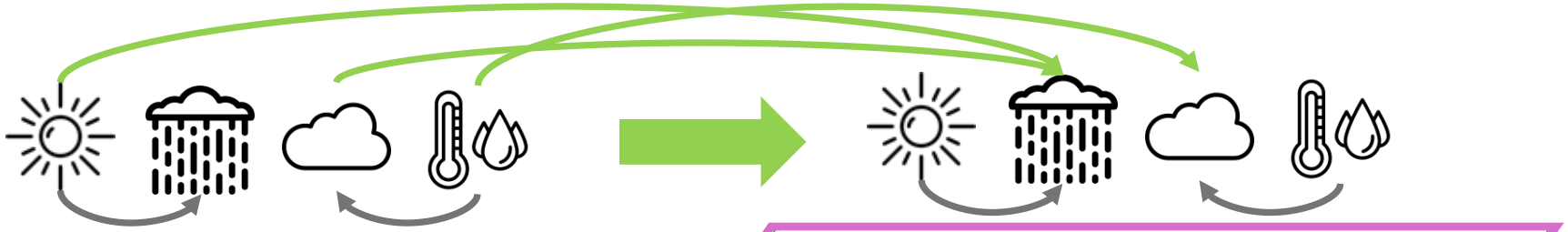
Dataset	Length	CaDRe		TDRL		CARD		FITS		MICN		iTransformer		TimesNet		Autoformer		Timer-XL	
		MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓
CESM2	96	0.410	0.483	0.439	0.507	**0.409**	0.484	0.439	0.508	0.417	0.486	0.422	0.491	0.415	0.486	0.959	0.735	0.433	**0.425**
CESM2	192	**0.412**	**0.487**	0.440	0.508	0.422	0.493	0.447	0.515	1.559	0.984	0.425	0.495	0.417	0.497	1.574	0.972	0.454	0.524
CESM2	336	**0.413**	**0.485**	0.441	0.505	0.421	0.497	0.482	0.536	2.091	1.173	0.426	0.494	0.423	0.499	1.845	1.078	0.527	0.565
Weather	96	**0.157**	**0.203**	0.442	0.511	0.423	0.497	0.172	0.221	0.199	0.256	0.168	0.214	0.180	0.231	0.225	0.259	0.367	0.252
Weather	192	0.207	0.248	0.492	0.545	0.482	0.544	0.216	0.260	0.238	0.298	**0.193**	**0.241**	0.212	0.265	0.354	0.348	0.434	0.298
Weather	336	**0.270**	**0.314**	0.536	0.612	0.525	0.596	0.386	0.439	0.316	0.496	0.426	0.494	0.423	0.499	0.354	0.348	0.527	0.565
ERSST	96	**0.145**	0.268	0.187	0.268	0.197	0.273	0.539	0.297	0.726	0.765	0.247	0.264	0.432	0.508	0.953	0.272	0.163	**0.259**
ERSST	192	**0.208**	0.307	0.214	**0.293**	0.233	0.375	0.226	0.752	1.263	0.892	0.251	0.535	0.452	0.585	1.024	0.908	0.210	0.294
ERSST	336	**0.305**	0.361	0.462	0.388	0.487	0.484	0.439	0.535	1.173	1.172	0.305	0.659	0.581	0.607	1.387	1.353	0.352	**0.337**

Tab. Weather forecasting across 3 different datasets

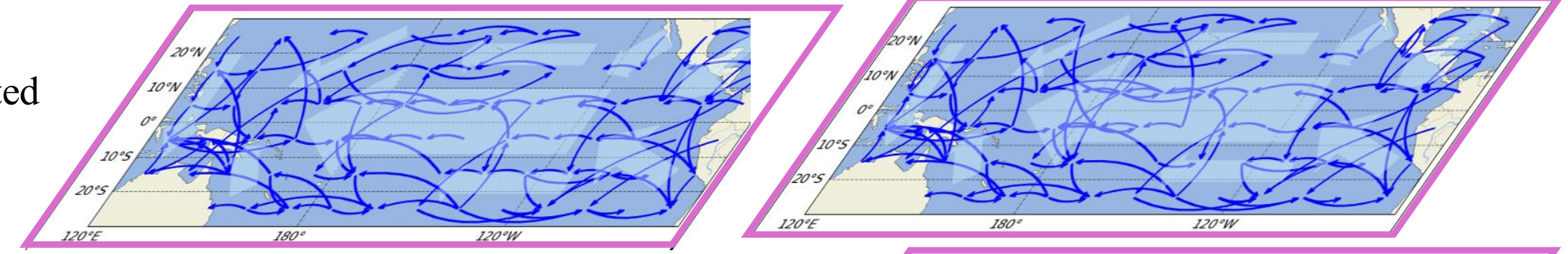
Experiment Results: Causal Learning on Climate System

- The overall wind trend in each map shows the consistency between **wind and causal relations**.
- **Scientific discoveries** by CaDRe are consistent with established scientific evidences.

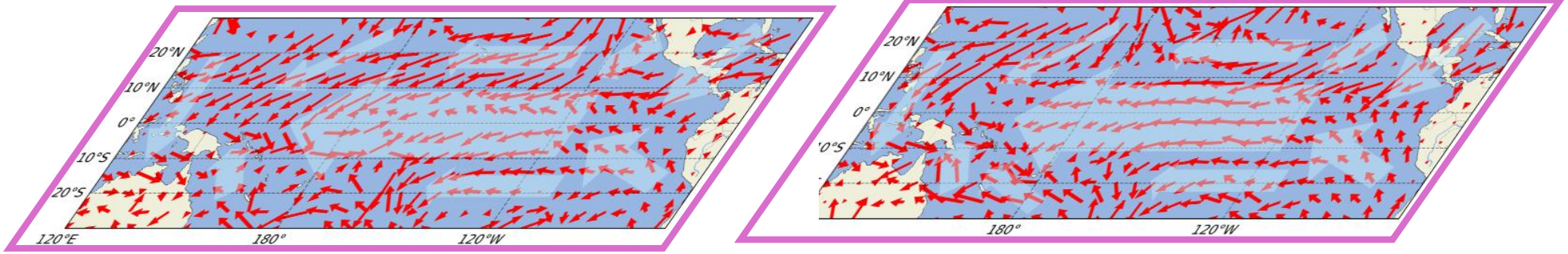
Estimated Latent Process



Estimated Causal Graph



Wind System



Limitations & Future Work

- Allow time-lagged causal relations over observed variables with 4-measurements (Hu & Shum, 2012)

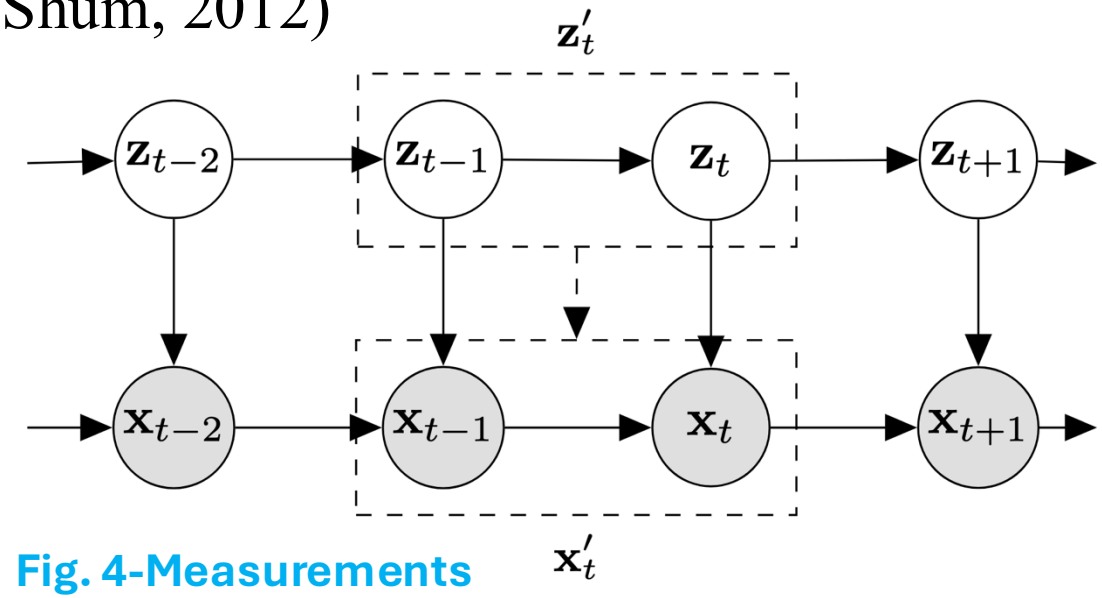


Fig. 4-Measurements

- A general state-space model for causal discovery

- Extend CaDRe to the climate foundational model

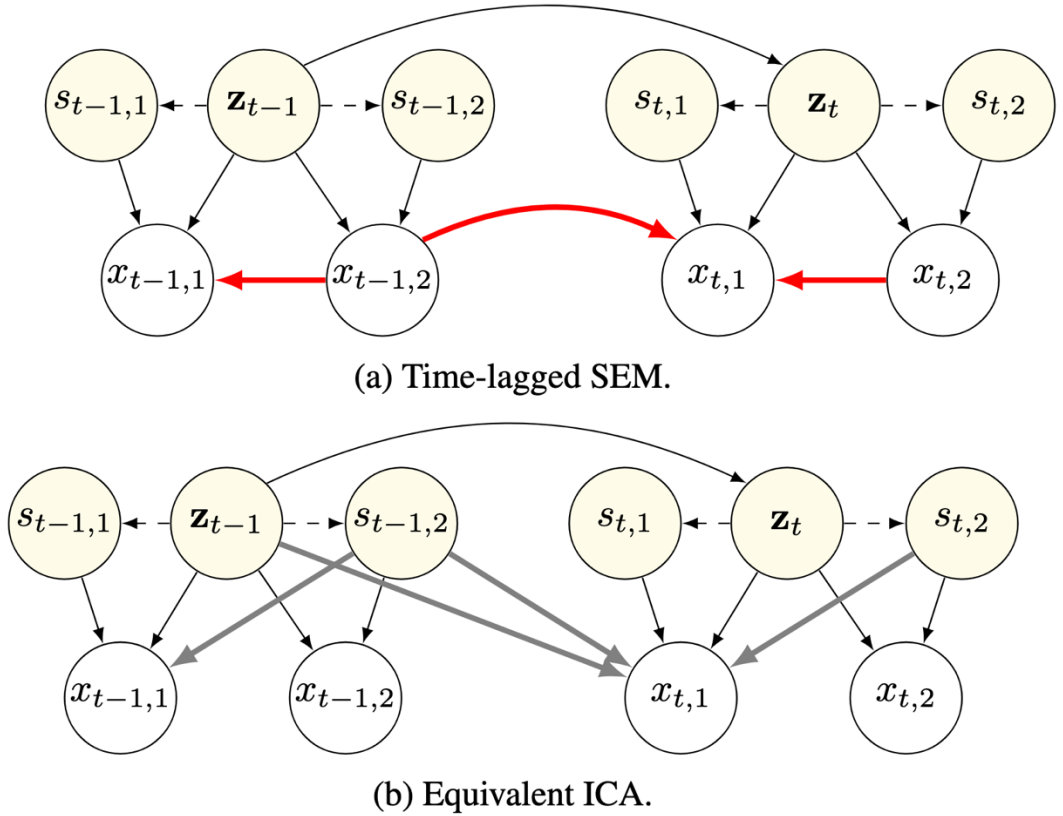


Fig. General State-space Model

Hu, Yingyao, and Matthew Shum. "Nonparametric identification of dynamic models with unobserved state variables." *Journal of Econometrics* 171.1 (2012): 32-44.

Thank You!

Minghao Fu
Ph.D. Student at UCSD