Towards Causal Representations of Climate Model Data

Julien Boussard[1,3*] **Chandni Nagda**[1,4*] **Julia Kaltenborn**[1,5]
Charlotte Emilie Elektra Lange[1,6] **Philippe Brouillard**[1,7]
Yaniv Gurwicz[2] **Peer Nowack**[8] **David Rolnick**[1,5,7]

[1]Mila –Quebec AI Institute [2]Intel Labs [3]Columbia University
[4]University of Illinois at Urbana-Champaign [5]McGill University [6]Osnabrück University
[7]Université de Montréal [8]Karlsruhe Institute of Technology *-equal contribution

Abstract

Climate models, such as Earth system models (ESMs), are crucial for simulating future climate change based on projected Shared Socioeconomic Pathways (SSP) greenhouse gas emissions scenarios. While ESMs are sophisticated and invaluable, machine learning-based emulators trained on existing simulation data can project additional climate scenarios much faster and are computationally efficient. However, they often lack generalizability and interpretability. This work delves into the potential of causal representation learning, specifically the *Causal Discovery with Single-parent Decoding* (CDSD) method, which could render climate model emulation efficient *and* interpretable. We evaluate CDSD on multiple climate datasets, focusing on emissions, temperature, and precipitation. Our findings shed light on the challenges, limitations, and promise of using CDSD as a stepping stone towards more interpretable and robust climate model emulation.

1 Introduction

Climate models are indispensable for simulating future climate scenarios based on Shared Socioeconomic Pathways (SSP) emissions. Earth system models (ESMs) are complex models grounded in systems of differential equations that capture a vast array of physical processes. They provide a comprehensive understanding of climate dynamics, but are computationally expensive, as even a single ESM run for one SSP requires about a year to run on a supercomputer (3). Recent advancements in data-driven models using machine learning (ML) present an opportunity to emulate climate projections more efficiently (14; 27). However, these climate model emulators often act as "black boxes", lacking interpretability crucial for climate science.

Causal methods enable the discovery and quantification of causal dependencies in observed data (22; 16), and have emerged as a valuable tool to improve our understanding of physical systems across various fields, including Earth Sciences (see (20)). In regards to climate modeling, causal methods can potentially bridge the gap between well defined, but compute-intensive ESMs and efficient, but "black-box" ML models. They could be used for A) causal evaluation of climate model emulators, by identifying causal dependencies within their projections and verifying that they correspond to known physical processes; B) climate emulation, by inferring causally-informed high-level representations underlying climate projections ; and C) causal hypothesis testing and attribution of climate change or extreme events. In particular for climate model emulation and the evaluation of those emulators, causal methods have not been used yet. In this work, we aim to investigate the potential of causal methods in the context of climate model emulators.

Previous work has leveraged causal methods in various forms to increase our understanding of climate data. The necessary foundation for quantifying causality from time-series, was laid by Granger

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023.

causality (8). It was used to infer causal links between CO_2 concentration and global temperature (25) and was later extended to deduce causal feedbacks in climate models (26). Building on this, approaches like PCMCI and PCMCI+ (19; 17; 21; 18) were developed to autonomously discover the causal graphs of observed systems, finding applications in climate science (15). More recent work such as Varimax-PCMCI (24) combines PCMCI with dimensionality reduction methods to derive causally-informed representations from observable climate data. However, since these methods do not scale well and become computationally inefficient for large and non-linear climate datasets, (2) introduced a differentiable causal discovery method called *Causal Discovery with Single-parent Decoding* (CDSD). CDSD uniquely learns both the latent representation from time-series and the causal graph over these latents simultaneously. Here, we are leveraging CDSD to uncover low-dimensional latent drivers that encapsulate temporal processes in climate model data.

Our primary objective is to harness the power of causal methods to make ML-based climate emulation more interpretable and trusted by experts. As a first step, we apply temporal causal representation learning, in the form of CDSD, to learn causally informed low-dimensional latents underlying the emissions, temperature, and precipitation time-series from ClimateSet (11). We show that CDSD is able to infer representations that match known physical processes, demonstrating that CDSD can be used to evaluate and validate climate models. We believe that our findings and proposed solutions pave the way for further development of causally informed climate model emulation techniques.

2 Causal Discovery with Single-Parent Decoding

CDSD considers a generative model where d_x-dimensional variables $\{\boldsymbol{x}^t\}_{t=1}^T$ are observed across T time steps. These observations, $\boldsymbol{x}^t$, are influenced by d_z-dimensional latent variables $\boldsymbol{z}^t$. For instance, $\boldsymbol{x}^t$ could represent climate measurements, while $\boldsymbol{z}^t$ might represent unknown regional climate trends.

The model considers a stationary time series of order τ over these latent variables. Binary matrices $\{G^k\}_{k=1}^{\tau}$ represent the causal relationships between latents at different time steps. Specifically, an element $G^k_{ij} = 1$ indicates that z_j^{t-k} is a causal parent of z_i^t, capturing the lagged causal relations between the time-steps $t-k$ and t. The adjacency matrix F delineates the causal connections between the latent variables $\boldsymbol{z}$ and the observed variables $\boldsymbol{x}$. Each observed variable x_i has at most one latent parent, adhering to the *single-parent decoding* structure. A high-level description of this model is provided here, with comprehensive details presented in Appendix A.

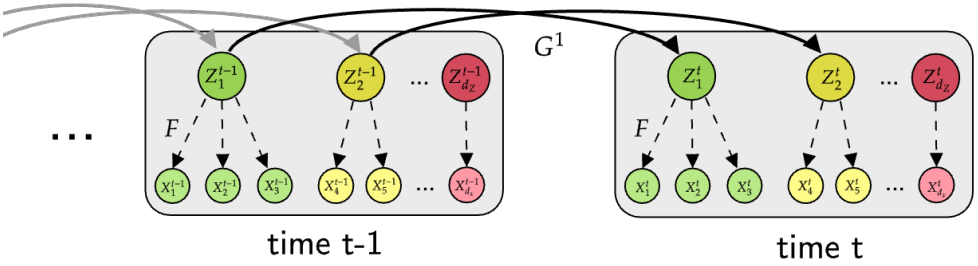

Figure 1: Generative model. Variables $\boldsymbol{z}$ are latent, and $\boldsymbol{x}$ are observable. G^k represents latent connections across different time lags, with the diagram only containing connections up to G^1. F connects latents to observables. Connections are illustrated only up to G^1, but CDSD leverages connections of higher order. *Figure reprinted with permission from (2).*

At any given time step t, the latents are assumed to be independent given their past, and each conditional is parameterized by a non-linear function g_j. h is chosen to be a Gaussian density function.

$$p(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1}, \ldots, \boldsymbol{z}^{t-\tau}) := \prod_{j=1}^{d_z} p(z_j^t \mid \boldsymbol{z}^{t-1}, \ldots, \boldsymbol{z}^{t-\tau}); \tag{1}$$

$$p(z_j^t \mid \boldsymbol{z}^{<t}) := h(z_j^t;\ g_j([G^1_{j:} \odot \boldsymbol{z}^{t-1}, \ldots, G^\tau_{j:} \odot \boldsymbol{z}^{t-\tau}])\,), \tag{2}$$

The observable variables x_j^t are assumed to be conditionally independent where $f_j : \mathbb{R} \to \mathbb{R}$, and $\boldsymbol{\sigma}^2 \in \mathbb{R}^{d_x}_{>0}$ are decoding functions:

$$p(x_j^t \mid \boldsymbol{z}^t_{pa_j^F}) := \mathcal{N}(x_j^t; f_j(\boldsymbol{z}^t_{pa_j^F}), \sigma_j^2), \tag{3}$$

The model's complete density is:

$$p(\boldsymbol{x}^{\leq T}, \boldsymbol{z}^{\leq T}) := \prod_{t=1}^{T} p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}) p(\boldsymbol{x}^t \mid \boldsymbol{z}^t). \tag{4}$$

Maximizing $p(\boldsymbol{x}^{\leq T}) = \int p(\boldsymbol{x}^{\leq T}, \boldsymbol{z}^{\leq T})\, d\boldsymbol{z}^{\leq T}$ unfortunately involves an intractable integral, hence the model is fit by maximizing an evidence lower bound (ELBO) (12; 7) for $p(\boldsymbol{x}^{\leq T})$. The variational approximation of the posterior $p(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T})$ is $q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T})$.

$$q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}) := \prod_{t=1}^{T} q(\boldsymbol{z}^t \mid \boldsymbol{x}^t); \quad q(\boldsymbol{z}^t \mid \boldsymbol{x}^t) := \mathcal{N}(\boldsymbol{z}^t; \tilde{\boldsymbol{f}}(\boldsymbol{x}^t), \mathrm{diag}(\tilde{\boldsymbol{\sigma}}^2)), \tag{5}$$

$$\log p(\boldsymbol{x}^{\leq T}) \geq \sum_{t=1}^{T} \Big[\mathbb{E}_{\boldsymbol{z}^t \sim q(\boldsymbol{z}^t \mid \boldsymbol{x}^t)} \left[\log p(\boldsymbol{x}^t \mid \boldsymbol{z}^t) \right] - \mathbb{E}_{\boldsymbol{z}^{<t} \sim q(\boldsymbol{z}^{<t} \mid \boldsymbol{x}^{<t})} \mathrm{KL} \left[q(\boldsymbol{z}^t \mid \boldsymbol{x}^t) \,||\, p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}) \right] \Big]. \tag{6}$$

The graph between the latent $\boldsymbol{z}$ and the observable $\boldsymbol{x}$ is parameterized using a weighted adjacency matrix W. To enforce the single-parent decoding, W is constrained to be non-negative with orthonormal columns. Neural networks are optionally used to parameterize encoding and decoding functions g_j, f_j, $\tilde{\boldsymbol{f}}$. The graphs G^k are sampled from $G^k_{ij} \sim Bernoulli(\sigma(\Gamma^k_{ij}))$, with Γ^k being learnable parameters. The objective is optimized using stochastic gradient descent, leveraging the Straight-Through Gumbel estimator (13; 10) and the reparameterization trick (12). See Appendix A for more details.

3 Results on Climate Data

Our experiments use emission, temperature, and precipitation data from ClimateSet (11), which extends ClimateBench (27), and is sourced from the Coupled Model Intercomparison Project Phase 6 (CMIP6) (6) and the Input Datasets for Model Intercomparison Projects (Input4MIPs) (4). From CMIP6, we selected the temperature and precipitation outputs of the Nor-ESM2 model, and from Input4MIPs the CO_2, CH_4, SO_2, and black carbon (BC) emission input data. For all input and output data the SSP2-4.5 scenario was chosen. For all variables, we consider the time-series spanning 2015 – 2100, with monthly frequency. Each time step corresponds to a 144×96 spatial global map (250 km resolution). For the experiments presented in Fig. 2, we ran CDSD on this data using a latent dimension $d_z = 50$, and inputs of time-length $\tau = 5$. In other words, we want to discover up to 50 latent variables, with causal links emerging over 5 months. Additional parameters and experimental setup are detailed in Appendix B.

Fig. 2(A) shows the clustering induced by CDSD for temperature across all grid point locations on the globe. Clusters represent grid points across the globe that have a common parent in the latent causal graph. These clusters can be interpreted as regions with similar climatic properties. Causal graphs over the latents represent the potential causal links between these spatial points across different time lags. Most clusters are found outside the tropics. This grouping can be understood from the larger seasonal variation in temperatures at higher latitudes, again combined with also larger high-latitude warming under greenhouse gas forcing (e.g. polar amplification (5; 23)). There is a clear land-ocean separation in the northern hemisphere. The causal clusters and linkages uncovered by CDSD on temperature correspond to the different climatic zones on Earth and captures their seasonal trends. According to our analysis, CDSD does not find meaningful causal connections between the recovered clusters and captures only seasonal variations but no forced trends. One possible explanation is that seasonal variations are magnitudes larger than forced trends, impeding the recovery of the trends relevant to climate modeling questions.

Similarly, Fig. 2(B) shows the clustering induced by CDSD when ran on precipitation data. Clusters appear spread out and less compact than those obtained with temperature data. This can be attributed to the nature of precipitation data, which is more patterned and regional (relatively wet and dry regions occur both at high and low latitudes). The clusters also highlight several semi-permanent large-scale high and low pressure systems, reflecting regions of low and high total rainfall respectively. The "dynamic" movement of air masses that are much more influential on precipitation than on temperature explains the elongated shape of many clusters. As weather systems move across ocean and land masses, there is not as clear a land-ocean separation as for temperature. Dry areas such as the Sahara desert, and rainforest areas, such as Amazon and Congo Basin, with high levels of average precipitation are visible.

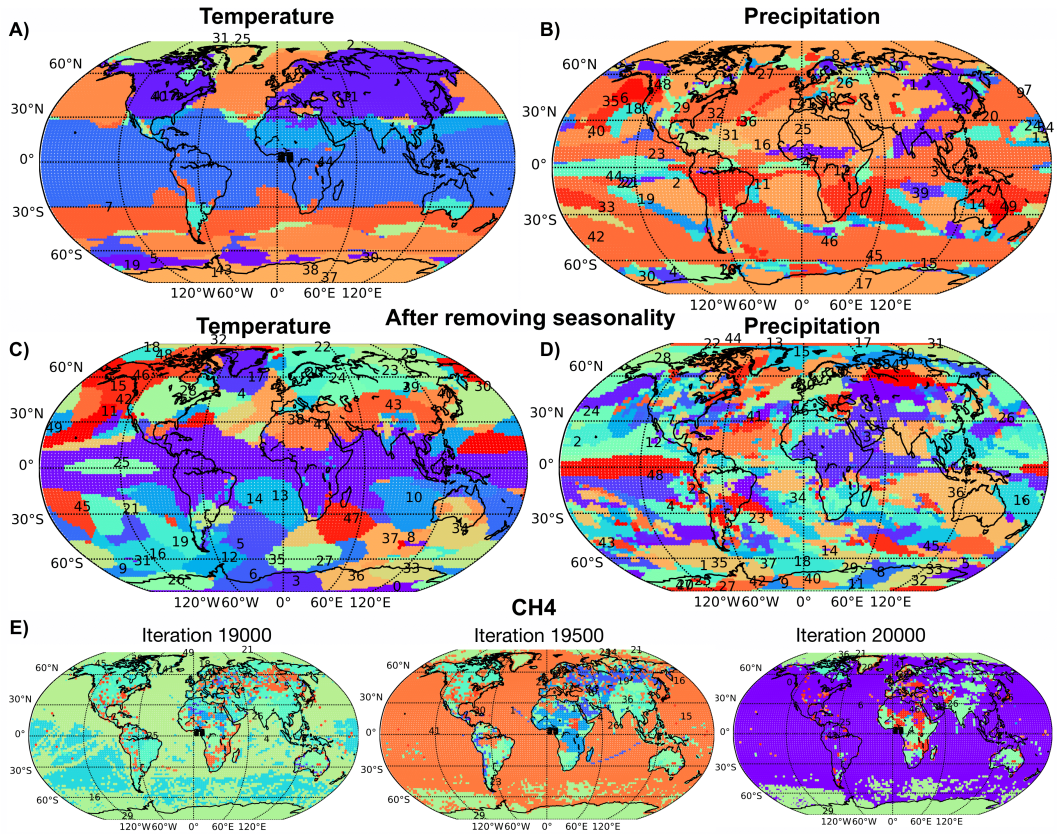

Figure 2: Segmentation of the Earth's surface according to the mapping between observations and latents learnt by CDSD for **(A)** CMIP6 temperature data of SSP2-4.5; **(B)** CMIP6 precipitation data of SSP2-4.5; **(C)** CMIP6 temperature and precipitation data of SSP2-4.5 after subtracting seasonal trends; and **(D)** CH_4 Input4MIPs data, at 3 different iterations after the loss plateaus. Each location is colored according to its parent in the causal representation.

Large-scale atmospheric circulation phenomena governing precipitation, such as convergence zones become visible, as well as clusters following tropical and extra-tropical cyclone patterns and their storm tracks. Clusters seem to broadly recover the Intertropical Convergence Zone (ITCZ) seasonal differences, as ITCZ is close to the equator during winter and moves above India and through China during the Northern Hemisphere's summer months.

To capture physical phenomena independent of seasonal variations, we perform "seasonality removal", and normalize the data by subtracting the monthly mean and dividing by the monthly standard deviation. The resulting clusters for temperature data, shown in Fig. 2(C) appear significantly different. Climatic zones - especially within tropical, subtropical and temperate zones - are now recovered by CDSD for temperature data. While there is still a dominant cluster along the equator, CDSD now finds a mid-Pacific cluster associated with the El Niño Southern Oscillation (ENSO), cluster 25, which corresponds to cluster 44 in (B). The zonal tropospheric circulation - see clusters surrounding the Antarctica in (A) and (B) - is more clearly patterned when removing seasonality. As the location and strength of those pressure cells is varying distinctly with seasons, their corresponding clusters might become more pronounced and cover more space when removing seasonality.

When seasonality is not removed from the precipitation data, the observed patterns closely align with sea level pressure maps, capturing broader climatic trends. However, by removing seasonality, the CDSD model reveals more localized climate phenomena. We postulate that in this more difficult regime, the model cannot rely on large seasonal rainfall differences to form clusters, and thus captures more granular detail. For instance, Antarctica is segmented into a greater number of clusters when seasonality is removed (Fig. 2 (D)). CDSD may be picking up on transient storm systems influenced by the phase behavior of the Southern Annular Mode (SAM). SAM is known for causing a poleward shift of storm tracks, as well as increased precipitation in the southern parts of Australia and South

America. Furthermore, rainforest clusters appear more detailed, with the separation between tropical and subtropical parts of the rainforests becoming more apparent. Southeast Alaska, the northern-most temperate zone of the Northern Hemisphere also appears separately as part of the purple cluster (24). Cluster 30, the light-green cluster over western-central Europe, likely corresponds to the storm tracks of eastwards-moving wet weather systems from the North Atlantic into Europe. Other phenomena are less clear. For example, clusters over Antarctica are very sensitive to parameterization: In some runs Antarctica's orography ("landscape") seems to influence local precipitation patterns strongly, while in others those patterns are less clear. Such phenomena need to be evaluated in further experiments.

Fig. 2(E) shows the clustering induced by CDSD ran on CH_4 emissions data, at different iterations before convergence. The model converges after 20000 iterations, according to the convergence criteria (detailed in Appendix B) of the optimized loss (Equation 8 in Appendix A). At each iteration shown in the figure, the loss is very close to the convergence loss, but the induced clustering is very different, showing that multiple representations correspond to similar objective values. CDSD is not able to robustly capture ships' CH_4 emissions, sometimes represented by clusters forming lines between different ports. It is not a dominant part of CH_4 emissions, but clearly shows that CDSD does not converge. We can see very different results in regions such as South-East Asia, America or oceans, with very different number of clusters being found at each iteration. It seems harder to find a stable causal representation underlying anthropogenic emission data than for physical climate variable data.

4 Discussion

CDSD is able to represent meaningful physical processes of temperature and precipitation measurements on the seasonal-to-century scale. We demonstrate results that coincide with known phenomena, such as regional temperature trends, ENSO, tropical and extra-tropical cyclone routes. We have highlighted the need for removing large seasonal variations that otherwise dominate forced trends. This approach could potentially remove the imprint of the seasonal cycle. To avoid this potential failure, one could standardize data with respect to the standard deviation and variable-dependent average over the complete time-series. Future work will explore modified approaches to distinguish forced and seasonal trends.

CDSD fails to stably represent emissions data. This makes sense, since forcing agent emissions are dominantly driven by anthropogenic effects and not natural physical processes. Human policy decisions and economic activities do not adhere to predictable temporal causal relationships and are thus not recoverable by CDSD. Furthermore, the lifetimes of different gases and aerosols range from weeks to hundreds of years, and impact the climate at various spatio-temporal resolutions. For example, carbon dioxide can persist in the atmosphere for thousands of years, leading to long-term cumulative warming effects. On the other hand, methane has a relatively short atmospheric lifetime of less than 12 years, but traps more heat (1). These discrepancies make for a particularly challenging task, as CDSD expects evenly sampled inputs with consistent time resolutions and expects causal temporal relationships to manifest within the relatively small time length τ. However, the number of parameters in the model scales linearly with τ; hence, increasing it beyond several years is computationally expensive and will make convergence difficult. Currently, we run CDSD using $\tau = 5$. For future work, we plan to use Transformers or Long Short-Term Memory to parameterize the transition functions g_j to handle larger values of τ. Also, as emissions are a cause of changes in observed climate variables, it might be possible to hard-code this causal relation and condition the learnt representation on the emissions. Such changes might allow CDSD to represent forced trends and improve climate projection.

5 Conclusion

Here, we demonstrated a first successful application of CDSD to investigate causal links in climate model data, and highlight future challenges in applying CDSD to emission data. For temperature and precipitation, the learned representations could be used to compare and evaluate different climate models and/or observations. Using CDSD in ML-based climate model emulators remains a challenge, albeit with the potential to render those emulators, for the first time, interpretable. Future work will consider crucial next steps (e.g., timescales of interest) towards such efficient and interpretable emulators, which are needed for the climate modeling community and ultimately to help inform policymakers.

References

[1] C. Adler, P. Wester, I. Bhatt, C. Huggel, G.E. Insarov, M.D. Morecroft, V. Muccione, and A. Prakash. *Cross-Chapter Paper 5: Mountains*, pages 2273–2318. Cambridge University Press, Cambridge, UK and New York, USA, 2022. ISBN 9781009325844. doi: 10.1017/9781009325844.022.2273.

[2] Anonymous. Causal Representation Learning in Temporal Data via Single-Parent Decoding. Submitted for review, 2024.

[3] Venkatramani Balaji, Eric Maisonnave, Niki Zadeh, Bryan N Lawrence, Joachim Biercamp, Uwe Fladrich, Giovanni Aloisio, Rusty Benson, Arnaud Caubel, Jeffrey Durachta, et al. Cpmip: measurements of real computational performance of earth system models in cmip6. *Geoscientific Model Development*, 10(1):19–34, 2017.

[4] Paul J Durack, Karl E Taylor, Veronika Eyring, Sasha K Ames, Charles Doutriaux, Tony Hoang, Denis Nadeau, Martina Stockhause, and Peter J Gleckler. input4mips: Making [cmip] model forcing more transparent. Technical report, Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), 2017.

[5] Mark R. England, Ian Eisenman, Nicholas J. Lutsko, and Till J. W. Wagner. The recent emergence of arctic amplification. *Geophysical Research Letters*, 48(15):e2021GL094086, 2021. doi: https://doi.org/10.1029/2021GL094086. URL `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021GL094086`. e2021GL094086 2021GL094086.

[6] Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5): 1937–1958, 2016.

[7] Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*, 2020.

[8] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/1912791`.

[9] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

[10] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[11] Julia Kaltenborn, Charlotte Emilie Elektra Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Jakob Runge, Peer Nowack, and David Rolnick. ClimateSet: A large-scale climate model dataset for machine learning. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. Accepted September 2023. To be published in December 2023.

[12] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[13] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[14] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *International Conference on Machine Learning (ICML)*, 2023.

[15] Peer Nowack, Jakob Runge, Veronika Eyring, and Joanna D Haigh. Causal networks for climate model evaluation and constrained projections. *Nature communications*, 11(1):1415, 2020.

[16] Raanan Y Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. From temporal to contemporaneous iterative causal discovery in the presence of latent confounders. *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[17] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947. PMLR, 2018.

[18] 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.

[19] Jakob Runge, Vladimir Petoukhov, Jonathan F Donges, Jaroslav Hlinka, Nikola Jajcay, Martin Vejmelka, David Hartman, Norbert Marwan, Milan Paluš, and Jürgen Kurths. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature communications*, 6(1): 1–10, 2015.

[20] 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):1–13, 2019.

[21] 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.

[22] Jakob Runge, Andreas Gerhardus, Gherardo Varando, Veronika Eyring, and Gustau Camps-Valls. Causal inference for time series. *Nature Reviews Earth & Environment*, 10:2553, 2023.

[23] D. M. Smith, J. A. Screen, C. Deser, J. Cohen, J. C. Fyfe, J. García-Serrano, T. Jung, V. Kattsov, D. Matei, R. Msadek, Y. Peings, M. Sigmond, J. Ukita, J.-H. Yoon, and X. Zhang. The polar amplification model intercomparison project (pamip) contribution to cmip6: investigating the causes and consequences of polar amplification. *Geoscientific Model Development*, 12(3):1139–1164, 2019. doi: 10.5194/gmd-12-1139-2019. URL `https://gmd.copernicus.org/articles/12/1139/2019/`.

[24] Xavier-Andoni Tibau, Christian Reimers, Andreas Gerhardus, Joachim Denzler, Veronika Eyring, and Jakob Runge. A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections. *Environmental Data Science*, 1:e12, 2022.

[25] Umberto Triacca. Is granger causality analysis appropriate to investigate the relationship between atmospheric concentration of carbon dioxide and global surface air temperature? *Theoretical and Applied Climatology*, 81:133–135, 07 2005. doi: 10.1007/s00704-004-0112-1.

[26] Egbert H. van Nes, Marten Scheffer, Victor Brovkin, Timothy M. Lenton, Hao Ye, Ethan Deyle, and George Sugihara. Causal feedbacks in climate change. *Nature Climate Change*, 5 (5):445–448, 2015. URL `https://EconPapers.repec.org/RePEc:nat:natcli:v:5:y:2015:i:5:d:10.1038_nclimate2568`.

[27] Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10):e2021MS002954, 2022.

A Inference with CDSD: Objective and Optimization

In this section, we present how inference and optimization is carried out when using CDSD (2).

Continuous optimization. The graphs G^k are learnt via continuous optimization. They are sampled from distributions parameterized by $\Gamma^k \in \mathbb{R}^{d_z \times d_z}$ that are learnable parameters. Specifically, $G^k_{ij} \sim Bernoulli(\sigma(\Gamma^k_{ij}))$, where $\sigma(\cdot)$ is the sigmoid function. This results in the following constrained optimization problem, with $\boldsymbol{\phi}$ denoting the parameters of all neural networks (r_j, g_j, $\tilde{\boldsymbol{f}}$) and the learnable variance terms at Equations 3 and 5:

$$\begin{aligned} &\max_{W,\Gamma,\boldsymbol{\phi}} \mathbb{E}_{G\sim\sigma(\Gamma)}\left[\mathbb{E}_{\boldsymbol{x}}\left[\mathcal{L}_{\boldsymbol{x}}(W,\Gamma,\boldsymbol{\phi})\right]\right] - \lambda_s||\sigma(\Gamma)||_1 \\ &\text{s.t. } W \text{ is orthogonal and non-negative,} \end{aligned} \tag{7}$$

$\mathcal{L}_{\boldsymbol{x}}$ is the ELBO corresponding to the right-hand side term in Equation 6 and $\lambda_s > 0$ a coefficient for the regularisation of the graph sparsity. The non-negativity of W is enforced using the projected gradient on $\mathbb{R}_{\geq 0}$, and its orthogonality enforced using the following constraint:

$$h(W) := W^T W - I_{d_z} \, .$$

Thsi results in the final constrained optimization problem, relaxed using the *augmented Lagrangian method* (ALM):

$$\max_{W,\Gamma,\boldsymbol{\phi}} \mathbb{E}_{G\sim\sigma(\Gamma)}\left[\mathbb{E}_{\boldsymbol{x}}[\mathcal{L}_{\boldsymbol{x}}(W,\Gamma,\boldsymbol{\phi})]\right] - \lambda_s||\sigma(\Gamma)||_1 - \text{Tr}\left(\lambda_W^T h(W)\right) - \frac{\mu_W}{2}||h(W)||_2^2, \tag{8}$$

where $\lambda_W \in \mathbb{R}^{d_z \times d_z}$ and $\mu_W \in \mathbb{R}_{>0}$ are the coefficients of the ALM.

This objective is optimized using stochastic gradient descent. The gradients w.r.t. the parameters Γ are estimated using the Straight-Through Gumbel estimator (13; 10). The ELBO is optimized following the classical VAE models (12), by using the reparametrization trick and a closed-form expression for the KL divergence term since both $q(\boldsymbol{z}^t \mid \boldsymbol{x}^t)$ and $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t})$ are multivariate Gaussians. The graphs G and the matrix W are thus learnt end-to-end.

B Detailed Parameters and Experimental Setup

We train our models on our internal cluster and use a single Nvidia-RTX8000 with 32GB of RAM for each run.

Reusing the default parameters of CDSD detailed in (2), we use leaky-ReLU as activation functions for all neural networks. For the neural networks g_j fitting the non-linear dynamic, we used Multi-layer perceptrons (MLPs) with 2 hidden layers and 8 hidden units. For the neural network r_j fitting the non-linear encoding, we use a single neural network that receives as input the masked $W\boldsymbol{z}^t$ and an embedding (dimension 10) of the index $s(j)$ is concatenated to the input. This neural network has 2 hidden layers and 32 hidden units. Furthermore, we use the optimizer RMSProp (9).

As we encountered problems with convergence when running experiments on emissions data, we performed a hyperparameter search for learning rate and batch size. For all experiments, we tried learning rates among $\{1e-2, 1e-3, 1e-4\}$, and batch sizes among $\{64, 128, 256, 512\}$. We also tried learning rate decay, without better success. For experiments reported in Fig. 2, we used a learning rate of $1e-3$ and batch size of 64. For all runs, the data is divided with a split ratio of 0.9 for the training set and 0.1 for the validation set. Models are trained until convergence, determined once the validation loss has not improved in 1000 iterations. As recommended by (2), we tried multiple values for the regularization coefficient enforcing graph sparsity of CDSD $\{10, 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$.

C Additional Experiments

To validate our hypothesis and get a better understanding of what needs to be done to use causal representation learning models for climate emulation, we conducted multiple experiments using different inputs,

data preprocessing and parameters. We report a list of experiments, along with the insight we gained from them. All experiments were conducted with the experimental setting described in Appendix B.

We tried using latents of dimension $50, 100$ and 200, but the results did not differ a lot among each other, although training slowed down with increasing latent dimension. The number of latents was particularly relevant when running CDSD on precipitation data, as we expected that the distinct regions within shared precipitation clusters would eventually separate. However, this was not the case. For future work, we suggest to implement a constraint enforcing the connectivity of the clusters in order to represent causal connections between specific regions of the globe more explicitely.

As mentioned in the main text, we ran CDSD on the CO_2, SO_2, and black carbon (BC) emission input data from Input4MIPs, on SSP2-4.5. SO_2 and BC, two aerosols, have a lifetime of less than 1 month whereas CO_2 is cumulative in nature and lasts more than 300 years in the atmosphere (given a working carboncycle - otherwise it technically stays forever). These processes, along with CH_4 that has a lifetime of approximately 10 years, interact on very different resolutions. We plan to run CDSD on carbon monoxide (CO) from Input4MIPs, as this gas has a lifetime of $2 - -3$ months, corresponding to the resolution of the input data.

To check if CH_4 and CO_2 could be represented when using lower time-resolution, we aggregated the monthly data to create annual data, and used $\tau = 5$ as well as $\tau = 12$ to capture longer-term causal connections. However, this did not solve the convergence issue. It is possible that the number of data points is now too low, as, after aggregating the data, the number of training points is reduced by 12. We also tried to remove seasonality, by standardizing the data of each of the 12 months over different years, hoping to capture non-seasonal causal links. For all gases, we tried inputting time-series of multiple resolution, 1, 3, 6 and 12 months together in order to learn representations that are invariant to time-resolution. For all these experiments, the model did not converge to a single causal representation.

An additional difficulty might arise from the high spatial resolution of Input4MIPs data. Areas of high natural CH_4 emissions (e.g. wetlands) can be positioned naturally next to regions of overall low emissions or high anthropogenic emissions (e.g. cities, areas of fracking etc.). Therefore, spatial homogeneity is not to be expected and may lead to unstable behaviour. Spatially aggregating the data to get coarser resolution and spatial homogeneity may lead to more stable behavior, although we might lose some information (such as anthropogenic vs. natural emissions). We plan to run CDSD using different input spatial resolution, and study the behavior of CDSD further.