

---

# Learning Hierarchical Causal Representations of the Effects of Forcings on Temperature in Climate Models

---

**Shan Zhao**

Technical University of Munich  
Munich, Germany

**Ilija Trajkovic**

Karlsruhe Institute of Technology  
Karlsruhe, Germany

**Julia Kaltenborn**

McGill University & Mila  
Montreal, Canada

**Yaniv Gurwicz**

Intel Labs  
Israel

**Peer Nowack**

Karlsruhe Institute of Technology  
Karlsruhe, Germany

**David Rolnick**

McGill University & Mila  
Montreal, Canada

**Julien Boussard**

McGill University & Mila  
Montreal, Canada

## Abstract

Machine learning (ML) emulators provide a fast and cost-effective method to simulate climate change scenarios after being trained on Earth System Models projections. However, the black-box nature of those data-driven approaches limit the usability and trustworthiness of their outputs and in particular their use as causal attribution tools. Here, we develop a hierarchical causal representation learning framework applied to sea surface temperature fields from a state-of-the-art global climate model. As a key advance over previous work, our framework explicitly models both atmospheric dynamical interactions arising from internal climate variability and forced responses due to changes in atmospheric greenhouse gas and aerosol concentrations. When trained on future climate change scenarios, our method accurately predicts the long-term global mean and regional temperature evolution and shows physically realistic responses to perturbations in greenhouse gas and aerosol concentrations when evaluated on unseen scenarios. Our results underline the potential of causal representation learning frameworks for advancing climate model emulation.

## 1 Introduction

Traditional climate models are based on a large set of differential equations that represent an enormous variety of processes and interactions driving the evolution of Earth’s climate. Solving these equations over long time horizons is highly computationally expensive and requires major approximations, including limited numerical grid resolutions and various parameterizations for processes such as cloud formation and convection [Balaji et al., 2017, Mansfield et al., 2020]. This has led to a rapidly growing interest in ML approaches to climate modelling, given fast ML inference and the possibility to train the resulting models not only on climate model data, but also directly on Earth observations [Watt-Meyer et al., 2023, Clyne et al., 2026].

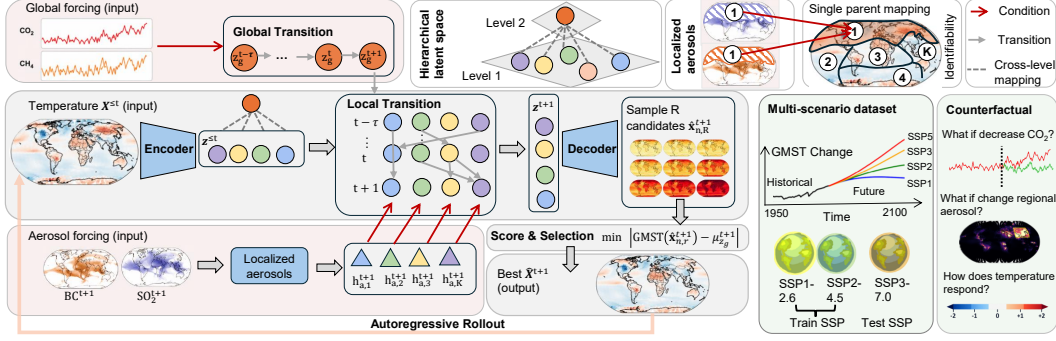


Figure 1: **Hierarchical climate emulator.** A global latent  $z_g$  captures climate responses to global forcings, while local latents  $z$  capture regional dynamics conditioned on spatially localized aerosols under the single-parent mapping. The model supports multi-scenario emulation and counterfactual perturbations of forcings.

While ML models can now outperform traditional, physics-based models at medium-range weather forecasting in several metrics [Bi et al., 2023, Lam et al., 2023, Lang et al., 2024, Bodnar et al., 2024], many of them have been demonstrated to show physically unrealistic behaviour when applied to long-term climate prediction [Chattopadhyay et al., 2024, Karlbauer et al., 2024]. Few ML models can address the task of climate projection – predicting the evolution of future climate given a scenario of greenhouse gas (GHG) and aerosol emissions – while also realistically simulating short-to-long-term weather and climate variability. The international community urgently needs the ability to project future climate under defined shared socioeconomic pathway (SSP) scenarios [Riahi et al., 2017], which provide different estimates of future emissions of anthropogenic forcings (GHG and aerosols).

The task of climate projection is extremely challenging given that anthropogenic climate change represents a long-term distribution shift in the presence of substantial natural internal variability across a large variety of spatial and temporal scales. Future observations come from a different distribution than past observations, while, ideally, a model should also be able to realistically capture short-term internal weather (hours-to weeks) to climate (years/decades) variability. ML models, which famously struggle with out-of-distribution (OOD) predictions [Ovadia et al., 2019], cannot be trusted in this context. Simple models, e.g. linear pattern scaling, cannot capture the complex processes underlying the evolution of the climate system, while increasing model complexity does not necessarily improve the emulation of the climate response to forcing scenarios, as expressive models may overfit correlations specific to training scenarios [Lütjens et al., 2025].

Causal representation learning has been proposed as a way to automatically learn a physically meaningful representation of the climate system, by learning the causal teleconnections between the observed variables of the modeled system [Brouillard et al., 2024, Boussard et al., 2023, Hickman et al., 2025]. The learned causal graph can thus represent both internal dynamics e.g. how the climatic variables naturally interact, and the response to external forcings. These score-based methods leverage differentiable constraints [Zheng et al., 2018, Brouillard et al., 2020] and have been applied to emulate the dynamics of sea-surface temperature in a pre-industrial control scenario e.g. without external forcings (PICABU, Hickman et al., 2025), successfully modeling the internal climate variability. In this paper, we learn to model the effect of various anthropogenic forcings on temperature, and explicitly separate the internal from the externally-driven variability of the climate system, at different spatial scales. We show that our model captures scenario-dependent warming trends and produces distinct temperature responses to GHG and regional aerosol perturbations. This work is a first step towards principled, causal, learned representations of the climate system, and the design of statistical attribution studies.

## 2 Methods

### 2.1 Model

PICABU [Hickman et al., 2025] jointly learns a low-dimensional latent representation  $\{z^t = (z_1, \dots, z_K)^t\}_{t \in \{t-\tau, \dots, t-1, t\}}$  of the high-dimensional climate states  $\{x\}_{t \in \{t-\tau, \dots, t-1, t\}}$  and their

dynamics, represented as sparse causal connections from  $\mathbf{z}^{<t}$  to  $\mathbf{z}^t$ . PICABU adopts the single-parent assumption, i.e., at a given timestep, each spatial grid point is mapped to a single latent  $z_k$ , resulting in  $K$  spatial clusters of grid points sharing similar climate patterns. This constrains the latent-to-climate mapping  $\mathbf{W}$  to satisfy the structural conditions required for identifiability. The model is trained by maximizing the evidence lower bound (ELBO) to maximize the probability of climate states (both in reconstructing  $x^{<t}$  and predicting  $x^t$ ) under constraint  $\mathcal{C}_{\text{single parent}}^{\{\lambda, \mu\}}$ , using the augmented Lagrangian method (ALM). Additionally, the temporal dependencies between local latents are constrained by  $\mathcal{C}_{\text{sparsity}}^{\{\lambda, \mu\}}$  to form a sparse transition graph.

**Hierarchical latent representation.** Building on this structure, we introduce a hierarchical latent representation and incorporate external climate forcings at different levels of the hierarchy. At the highest level, we introduce a global latent  $z_g \in \mathbb{R}$ , which yields the hierarchy  $z_g \rightarrow \mathbf{z} \rightarrow \mathbf{x}$  that retains the single-parent structure across both latent levels and thus preserves identifiability. Temporal dynamics are modeled at both levels:

$$\underbrace{(z_g^{<t}, \text{CO}_2^t, \text{CH}_4^t) \rightarrow z_g^t}_{\text{global transition}}, \quad \underbrace{(z_g^t, \mathbf{z}^{<t}, \text{BC}_k^t, \text{SO}_{2,k}^t) \rightarrow z_k^t}_{\text{local transition}}, \quad (1)$$

where global GHG forcings  $\text{CO}_2$ ,  $\text{CH}_4$  (carbon dioxide, methane) act through  $z_g$ , while aerosol forcings affect local latents.  $\text{SO}_{2,k}$  and  $\text{BC}_k$  denote the localized aerosol forcings (sulfur, black carbon) associated with  $z_k$ , as detailed in Eq. 2. To capture localized aerosol effects, we use  $\mathbf{W}$  to map aerosols into the same  $K$  clusters, and each climate latent is affected by aerosols only within its corresponding region. For  $a \in \{\text{BC}, \text{SO}_2\}$ , the localized forcing used to condition  $z_k$  is transformed by the nonlinear function  $f_{a,k}$ :

$$\mathbf{h}_{a,k}^t = f_{a,k}(\mathbf{W}_{:,k} \odot \mathbf{a}^t). \quad (2)$$

Therefore,  $\mathbf{W}$  jointly represents the spatial mapping to capture both climate variability and regional responses to aerosol forcing. We adapt the ELBO to the hierarchical model, now called  $\mathcal{L}_{\text{H-ELBO}}$  by factorizing the posterior and prior according to the latent hierarchy, decomposing the standard KL term into local- and global-level KL divergences. Details are provided in A.

**Global temperature constraints.** We introduce constraints on global mean surface temperature (GMST) at both the latent and observation levels. We encourage both the global latent and the global mean of the predicted temperature field to be close to the true GMST during training. With  $\mu_x^t$  the spatial mean of the predictive distribution  $p(x^t | \mathbf{z}^t)$  and  $\mu_g^t$  the mean of  $p(z_g^t | z_g^{<t})$ , we add the following loss function to our training objective:

$$\mathcal{L}_{\text{GMST}} = \|\mu_g^t - \text{GMST}^t\|_2^2 + \|\mu_x^t - \text{GMST}^t\|_2^2. \quad (3)$$

The overall core training objective is in Eq.4, where  $\mathcal{L}_{\text{aux}}$  comprises the probabilistic and spectral losses in PICABU [Hickman et al., 2025]:

$$\mathcal{L}_{\text{core}} := -\mathcal{L}_{\text{H-ELBO}} + \mathcal{C}_{\text{single parent}}^{\{\lambda, \mu\}} + \mathcal{C}_{\text{sparsity}}^{\{\lambda, \mu\}} + \lambda_g \mathcal{L}_{\text{GMST}} + \lambda_a \mathcal{L}_{\text{aux}} \quad (4)$$

## 2.2 Hierarchical autoregressive rollout

To generate long-term climate trajectories, we perform autoregressive rollouts by first predicting the global latent value at time  $T+1$  given the global forcing concentrations at  $T+1$ , and then predicting the local latents given the global latent value and local forcing concentrations, before decoding the temperature fields e.g. going from top to bottom in the hierarchy. We iterate this procedure to get full trajectories. Since our model is probabilistic, we generate  $N > 1$  trajectories. We initially draw  $N$  samples for the local latents, and, at each timestep, for each sample, we sample  $R$  candidates  $\mathbf{z}_{n,r}^{t+1} \sim p(\mathbf{z}^{t+1} | \mathbf{z}_n^{<t}, z_g^{t+1})$  before keeping the highest-scoring candidate, yielding  $N$  filtered trajectories for autoregressive propagation. The score associated with each sample is the L2 norm between the mean of the global latent  $\mu_g^{t+1}$  and the mean of the decoded temperatures. Thus, we ensure that the values of the local latents lead to a global mean temperature that is close to the global latent value, which itself is trained to approximate the true GMST. This is an extension of the Bayesian filtering technique used in PICABU, where we use our own predictions of the GMST rather than assuming a fixed spatial spectrum of the temperature field.

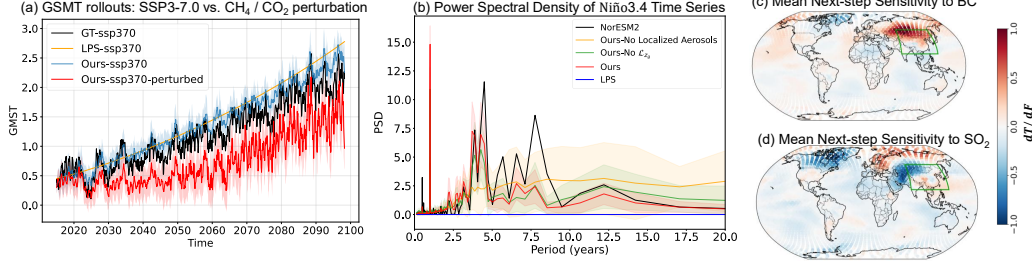


Figure 2: **Long-term climate rollouts and forcing perturbations.** (a) Our GMST rollouts under SSP3-7.0 (blue) compared with NorESM2 (black) and Linear Pattern Scaling (LPS) prediction (orange). Fixing  $\text{CO}_2$  and  $\text{CH}_4$  at their 2015 levels (red) suppresses the long-term increase in GMST. (b) Power spectral density of the Niño3.4 index under SSP3-7.0. The model ensemble mean (red) reproduces the characteristic Niño3.4 spectral peak. Shading in (a,b) denotes the std. across 50 trajectories. Details of the ablations and comparison to LPS are in C.2. (c,d) Next-step temperature responses to regional BC and  $\text{SO}_2$  perturbations over East Asia, showing warming and cooling responses, respectively. Green boxes indicate the perturbed region.

### 3 Experiments and results

#### 3.1 Experiments

We train and evaluate our model on monthly data from NorESM2-LM [Seland et al., 2020] simulations. It is trained to predict the monthly surface temperature from previous months’ surface temperatures and  $\text{CO}_2$ ,  $\text{CH}_4$ , BC, and  $\text{SO}_2$  concentrations at the current month, as given by the SSP scenarios. We train on the historical period, SSP1-2.6, and SSP2-4.5, and test on the OOD SSP3-7.0 scenario.  $\text{CO}_2$  and  $\text{CH}_4$  are included as global GHG forcings and BC and  $\text{SO}_2$  as spatially varying aerosol forcings. Temperature is deseasonalized using the pre-industrial monthly climatology, without forced trend, while all variables are standardized using historical statistics. We use  $K = 48$  local latents, one global latent,  $d_x = 3072$  grid points on a HEALPix grid (equal area grid, [Gorski et al., 2005]), and  $\tau = 5$  months. Rollouts use  $N = 50$  particles with  $R = 10$  candidates each. Full details are in B.

#### 3.2 Results

We evaluate long-term climate emulation using 85-year autoregressive rollouts from January 2015. Despite being trained only on SSP1-2.6 and SSP2-4.5, the model captures the stronger warming under the held-out SSP3-7.0 scenario (Fig. 2a). While still overestimating the overall warming trend, ours achieves the closer mean to NorESM2 for both GMST (1.41 vs. 1.24) and Niño3.4 (1.47 vs. 1.37) compared with baseline. The model also captures the characteristic Niño3.4 periodicity of approximately 2.5–5 years and the annual peak corresponding to seasonal variations (Fig. 2b). Additional rollout analyses are provided in C.1.

We then evaluate temperature responses to forcing perturbations. Fixing  $\text{CO}_2$  and  $\text{CH}_4$  at their 2015 levels suppresses the long-term increase in  $z_g$  (Fig. 17), accompanied by reduced GMST warming (Fig. 2a). The perturbed trajectories also exhibit a larger ensemble spread, indicating increased uncertainty under the modified forcing conditions. Regional aerosol perturbations produce localized, forcing-specific responses (Fig. 2c,d): increasing BC over East Asia induces temperature increases, whereas increasing  $\text{SO}_2$  produces temperature decreases, consistent with their dominant warming and cooling radiative effects, respectively Ramanathan and Feng [2009]. Additional forcing–response analyses are provided in C.3.

### 4 Discussion

Our hierarchical causal representation learning model is able to accurately predict the GMST from GHG and aerosol concentrations in one moderate extrapolation scenario. Furthermore, it allows us to disentangle local from global effects, as we explicitly condition local and global representations on forcings of different spatial scales. It is also able to distinguish the effects of two different aerosol concentrations, BC and  $\text{SO}_2$ , over East Asia. However, the current formulation does not guarantee

separation of individual forcings within the same hierarchical level, such as  $\text{CO}_2$  and  $\text{CH}_4$  at the global level. Future work will incorporate causal modules to better identify the effects of individual forcings within each level [Wang and Blei, 2019, Almodóvar et al., 2025].

The evaluation of our model is limited to realistic representations of the forced response and ENSO power spectra within a climate model, as there is no ground truth causal graph of, e.g. atmospheric dynamical interactions. Although there exist benchmarks for evaluating causal discovery [Herdeanu et al., 2025] and causal representation learning (SAVAR, Tibau et al., 2022), none of these benchmarks includes external forcings. In future work, we will extend SAVAR with external forcings to evaluate our model in a controlled setting, and benchmark it against a set of existing methods. Moreover, we will include more climatic variables, such as sea level pressure or precipitation, to give a more complete picture of the causal connections between forcings and climate variables. We will also evaluate our method on additional climate models (e.g. CESM2) and more challenging scenarios (e.g.  $4\times\text{CO}_2$ , SSP5-8.5). Finally, we hope to apply this work to observed extreme events, and perform a proper attribution study in a real-world use case.

## References

- Alejandro Almodóvar, Adrián Javaloy, Juan Parras, Santiago Zazo, and Isabel Valera. Decaflo: A deconfounding causal generative model, 2025. URL <https://arxiv.org/abs/2503.15114>.
- Venkatramani Balaji, Eric Maisonnave, Niki Zadeh, Bryan N. Lawrence, Joachim Biercamp, Uwe Fladrich, Giovanni Aloisio, Rusty Benson, Arnaud Caubel, Jeffrey Durachta, Marie-Alice Foujols, Grenville Lister, Silvia Mocavero, Seth Underwood, and Garrett Wright. CPMIP: measurements of real computational performance of Earth system models in CMIP6. *Geoscientific Model Development*, 10(1):19–34, January 2017. ISSN 1991-959X. doi: 10.5194/gmd-10-19-2017. URL <https://gmd.copernicus.org/articles/10/19/2017/>. Publisher: Copernicus GmbH.
- Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06185-3. URL <https://www.nature.com/articles/s41586-023-06185-3>. Publisher: Nature Publishing Group.
- Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Vaughan, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Tham-biratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A Foundation Model for the Earth System, November 2024. URL <http://arxiv.org/abs/2405.13063>. arXiv:2405.13063 [physics].
- Julien Boussard, Chandni Nagda, Julia Kaltenborn, Charlotte Emilie Elektra Lange, Philippe Brouillard, Yaniv Gurwicz, Peer Nowack, and David Rolnick. Towards causal representations of climate model data. *arXiv preprint arXiv:2312.02858*, 2023.
- Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable Causal Discovery from Interventional Data. In *Advances in Neural Information Processing Systems*, volume 33, pages 21865–21877. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/f8b7aa3a0d349d9562b424160ad18612-Abstract.html>.
- 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, October 2024. URL <http://arxiv.org/abs/2410.07013>. arXiv:2410.07013 [cs].
- Ashesh Chattopadhyay, Y. Qiang Sun, and Pedram Hassanzadeh. Challenges of learning multi-scale dynamics with AI weather models: Implications for stability and one solution, December 2024. URL <http://arxiv.org/abs/2304.07029>. arXiv:2304.07029 [physics].
- Graham Clyne, Guillaume Couairon, Guillaume Gastineau, Claire Monteleoni, and Anastase Charantonis. Archesclimate: Probabilistic decadal ensemble generation with flow matching, 2026. URL <https://arxiv.org/abs/2509.15942>.
- K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelmann. Healpix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2):759–771, April 2005. ISSN 1538-4357. doi: 10.1086/427976. URL <http://dx.doi.org/10.1086/427976>.
- Benjamin Herdeanu, Juan Nathaniel, Carla Roesch, Jatan Buch, Gregor Ramien, Johannes Haux, and Pierre Gentine. Causaldynamics: A large-scale benchmark for structural discovery of dynamical causal models, 2025. URL <https://arxiv.org/abs/2505.16620>.
- Sebastian Hickman, Ilija Trajkovic, Julia Kaltenborn, Francis Pelletier, Alex Archibald, Yaniv Gurwicz, Peer Nowack, David Rolnick, and Julien Boussard. Causal climate emulation with bayesian filtering. 2025. URL <https://arxiv.org/abs/2506.09891>.
- Matthias Karlbauer, Nathaniel Cresswell-Clay, Dale R. Durran, Raul A. Moreno, Thorsten Kurth, Boris Bonev, Noah Brenowitz, and Martin V. Butz. Advancing Parsimonious Deep Learning Weather Prediction Using the HEALPix Mesh. *Journal of Advances in Modeling Earth Systems*, 16(8):e2023MS004021, 2024. ISSN 1942-2466. doi: 10.1029/2023MS004021.

- URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2023MS004021>. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1029/2023MS004021>.
- Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Meroze, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. GraphCast: Learning skillful medium-range global weather forecasting, August 2023. URL <http://arxiv.org/abs/2212.12794>. arXiv:2212.12794 [cs].
- Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Dueben, Andrew Brown, Florian Pappenberger, and Florence Rabier. Aifs – ecmwf’s data-driven forecasting system, 2024. URL <https://arxiv.org/abs/2406.01465>.
- Björn Lütjens, Raffaele Ferrari, Duncan Watson-Parris, and Noelle E. Selin. The impact of internal variability on benchmarking deep learning climate emulators. *Journal of Advances in Modeling Earth Systems*, 17(8):e2024MS004619, 2025. doi: <https://doi.org/10.1029/2024MS004619>. URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024MS004619>. e2024MS004619 2024MS004619.
- L. A. Mansfield, P. J. Nowack, M. Kasoar, R. G. Everitt, W. J. Collins, and A. Voulgarakis. Predicting global patterns of long-term climate change from short-term simulations using machine learning. *npj Climate and Atmospheric Science*, 3(1):1–9, November 2020. ISSN 2397-3722. doi: 10.1038/s41612-020-00148-5. URL <https://www.nature.com/articles/s41612-020-00148-5>. Publisher: Nature Publishing Group.
- Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.
- Veerabhadran Ramanathan and Yan Feng. Air pollution, greenhouse gases and climate change: Global and regional perspectives. *Atmospheric environment*, 43(1):37–50, 2009.
- Keywan Riahi, Detlef P. van Vuuren, Elmar Kriegler, Jae Edmonds, Brian C. O’Neill, Shinichiro Fujimori, Nico Bauer, Katherine Calvin, Rob Dellink, Oliver Fricko, Wolfgang Lutz, Alexander Popp, Jesus Crespo Cuaresma, Samir KC, Marian Leimbach, Leiwen Jiang, Tom Kram, Shilpa Rao, Johannes Emmerling, Kristie Ebi, Tomoko Hasegawa, Petr Havlik, Florian Humpenöder, Lara Aleluia Da Silva, Steve Smith, Elke Stehfest, Valentina Bosetti, Jiyong Eom, David Gernaat, Toshihiko Masui, Joeri Rogelj, Jessica Streffer, Laurent Drouet, Volker Krey, Gunnar Luderer, Mathijs Harmsen, Kiyoshi Takahashi, Lavinia Baumstark, Jonathan C. Doelman, Mikiko Kainuma, Zbigniew Klimont, Giacomo Marangoni, Hermann Lotze-Campen, Michael Obersteiner, Andrzej Tabeau, and Massimo Tavoni. The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, 42: 153–168, 2017. ISSN 0959-3780. doi: <https://doi.org/10.1016/j.gloenvcha.2016.05.009>. URL <https://www.sciencedirect.com/science/article/pii/S0959378016300681>.
- Øyvind Seland, Mats Bentsen, Dirk Olivié, Thomas Toniazzo, Ada Gjermundsen, Lise Seland Graff, Jens Boldingh Debernard, Alok Kumar Gupta, Yan-Chun He, Alf Kirkevåg, et al. Overview of the norwegian earth system model (noresm2) and key climate response of cmip6 deck, historical, and scenario simulations, *geosci. model dev.*, 13, 6165–6200. 2020.
- 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, January 2022. ISSN 2634-4602. doi: 10.1017/eds.2022.11.
- Yixin Wang and David M. Blei. The blessings of multiple causes, 2019. URL <https://arxiv.org/abs/1805.06826>.

Oliver Watt-Meyer, Gideon Dresdner, Jeremy McGibbon, Spencer K. Clark, Brian Henn, James Duncan, Noah D. Brenowitz, Karthik Kashinath, Michael S. Pritchard, Boris Bonev, Matthew E. Peters, and Christopher S. Bretherton. Ace: A fast, skillful learned global atmospheric model for climate prediction, 2023. URL <https://arxiv.org/abs/2310.02074>.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL <https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html>.

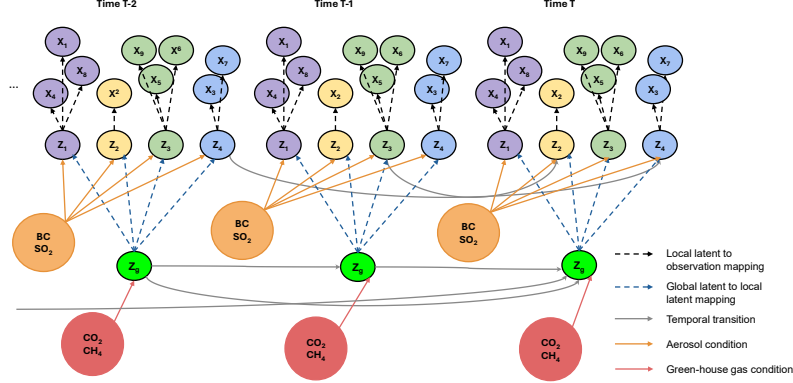


Figure 3: **Overview of the hierarchical generative model.** At each timestep, greenhouse-gas forcings ( $\text{CO}_2$  and  $\text{CH}_4$ ) condition the global latent  $z_g$ , whose temporal evolution depends on its previous states. The resulting global state conditions the local latents  $\mathbf{z}$ , which additionally depend on their sparse temporal parents and spatially localized aerosol forcings (BC and  $\text{SO}_2$ ). Aerosol forcing is assigned to each local latent according to the learned latent-to-observation mapping, while each observation is generated from its associated local latent under the single-parent structure. This hierarchy enables global greenhouse-gas responses and regional aerosol effects to enter the climate dynamics through different spatial levels.

## A Methods

### A.1 Hierarchical Generative Model

Let  $\mathbf{x}^t \in \mathbb{R}^{d_x}$  denote the spatial temperature field at time  $t$ . We represent the climate state by  $K$  local latent variables  $\mathbf{z}^t = (z_1^t, \dots, z_K^t)$  and a global climate latent  $z_g^t$ . The local climate representation is inferred from the temperature field, followed by the global representation,

$$q(\mathbf{z}^t, z_g^t | \mathbf{x}^t) = q_\phi(\mathbf{z}^t | \mathbf{x}^t) q_\psi(z_g^t | \mathbf{z}^t). \quad (5)$$

We model the global latent dynamics conditioned on greenhouse-gas forcings as

$$p(z_g^t | z_g^{<t}, \text{CO}_2^t, \text{CH}_4^t), \quad (6)$$

whose parameters are given by

$$(\mu_g^t, \sigma_g^t) = f_g(z_g^{<t}, \text{CO}_2^t, \text{CH}_4^t). \quad (7)$$

Conditioned on the current global state, each local climate latent follows

$$p(z_k^t | \text{Pa}(z_k^t), z_g^t, \mathbf{a}_k^t), \quad (8)$$

where  $\text{Pa}(z_k^t) \subseteq \mathbf{z}^{<t}$  is determined by the transition graph under  $\mathcal{C}_{\text{sparsity}}^{\{\lambda, \mu\}}$  and  $\mathbf{a}_k^t$  denotes the localized aerosol forcing associated with latent  $k$ .

The local transition is then parameterized as

$$(\mu_k^t, \sigma_k^t) = g_k(z_g^t, \text{Pa}(z_k^t), \mathbf{h}_{\text{BC}, k}^t, \mathbf{h}_{\text{SO}_2, k}^t), \quad (9)$$

where  $\mathbf{h}_{\text{BC}, k}^t$  and  $\mathbf{h}_{\text{SO}_2, k}^t$  denote the localized BC and  $\text{SO}_2$  forcing representations in Eq. 2. The overall graphical model is illustrated in Fig. 3.

### A.2 Hierarchical ELBO Derivation

We extend the original variational objective to a two-level latent hierarchy. Starting from the sequential variational objective,

$$\log p(\mathbf{x}^{\leq T}) \geq \sum_{t=1}^T \left[ \mathbb{E}_{q(\mathbf{z}^t | \mathbf{x}^t)} [\log p(\mathbf{x}^t | \mathbf{z}^t)] - \mathbb{E}_{q(\mathbf{z}^{<t} | \mathbf{x}^{<t})} \text{KL}[q(\mathbf{z}^t | \mathbf{x}^t) \| p(\mathbf{z}^t | \mathbf{z}^{<t})] \right]. \quad (10)$$

Let  $\mathbf{z}_1^t$  denote the local climate latents and  $z_2^t$  the high-level global latent. The hierarchical posterior is factorized as

$$q(\mathbf{z}_1^t, z_2^t | \mathbf{x}^t) = q(\mathbf{z}_1^t | \mathbf{x}^t)q(z_2^t | \mathbf{z}_1^t, \mathbf{x}^t) = q(\mathbf{z}_1^t | \mathbf{x}^t)q(z_2^t | \mathbf{z}_1^t). \quad (11)$$

where the last equality follows from the dependency structure encoded in our graphical model, in which  $z_2^t$  is conditionally independent of  $\mathbf{x}^t$  given  $\mathbf{z}_1^t$ .

The reconstruction term (first term) becomes

$$\mathbb{E}_{q(z_2^t | \mathbf{z}_1^t)q(\mathbf{z}_1^t | \mathbf{x}^t)} [\log p(\mathbf{x}^t | \mathbf{z}_1^t)] = \mathbb{E}_{q(\mathbf{z}_1^t | \mathbf{x}^t)} [\log p(\mathbf{x}^t | \mathbf{z}_1^t)], \quad (12)$$

where the observations are decoded directly from the local latent representation  $\mathbf{z}_1^t$ .

The prior follows the two-level hierarchy

$$p(\mathbf{z}_1^t, z_2^t | \mathbf{z}_1^{<t}, z_2^{<t}) = p(z_2^t | \mathbf{z}_1^{<t}, z_2^{<t})p(\mathbf{z}_1^t | \mathbf{z}_1^{<t}, z_2^{<t}, z_2^t) = p(z_2^t | z_2^{<t})p(\mathbf{z}_1^t | \mathbf{z}_1^{<t}, z_2^t), \quad (13)$$

where the last equality follows from the conditional dependencies specified by our graphical model,

$$z_2^t \perp\!\!\!\perp \mathbf{z}_1^{<t} | z_2^{<t}, \quad z_1^t \perp\!\!\!\perp z_2^{<t} | (z_1^{<t}, z_2^t). \quad (14)$$

Accordingly, the KL divergence term (second term) becomes

$$\begin{aligned} & \mathbb{E}_{q(z_2^{<t} | \mathbf{z}_1^{<t})q(\mathbf{z}_1^{<t} | \mathbf{x}^{<t})} \text{KL} [q(z_2^t | \mathbf{z}_1^t)q(\mathbf{z}_1^t | \mathbf{x}^t) || p(z_2^t | \mathbf{z}_2^{<t})p(\mathbf{z}_1^t | \mathbf{z}_1^{<t}, z_2^t)] = \\ & \mathbb{E}_{q(z_2^{<t} | \mathbf{z}_1^{<t})q(\mathbf{z}_1^{<t} | \mathbf{x}^{<t})} (\text{KL} [q(z_2^t | \mathbf{z}_1^t) || p(z_2^t | \mathbf{z}_2^{<t})] + \text{KL} [q(\mathbf{z}_1^t | \mathbf{x}^t) || p(\mathbf{z}_1^t | \mathbf{z}_1^{<t}, z_2^t)]) \end{aligned}$$

Combining the reconstruction and the KL terms gives

$$\begin{aligned} \mathcal{L}_{\text{H-ELBO}} = \sum_{t=1}^T & \left[ \mathbb{E}_{q(\mathbf{z}_1^t | \mathbf{x}^t)} [\log p(\mathbf{x}^t | \mathbf{z}_1^t)] - \text{KL} [q(\mathbf{z}_1^t | \mathbf{x}^t) || p(\mathbf{z}_1^t | \mathbf{z}_1^{<t}, z_2^t)] \right. \\ & \left. - \text{KL} [q(z_2^t | \mathbf{z}_1^t) || p(z_2^t | z_2^{<t})] \right]. \quad (15) \end{aligned}$$

## B Hyperparameters

Table 1 summarizes the model architecture and training hyperparameters used for the reported experiments.

## C Results

### C.1 Additional rollout analysis

We provide additional analyses of the long-term rollouts, including the SSP forcings (Fig. 4b) and the spatial PSD of the time-averaged temperature fields (Fig. 5). We further report the GMST (Fig. 6) and Niño3.4 time series (Fig. 7), together with their temporal autocorrelations (Fig. 8 and Fig. 9) and PSDs (Fig. 10). GMST is computed by spatially averaging the temperature field, while Niño3.4 is an El Niño-Southern Oscillation (ENSO) index computed by averaging sea-surface temperature anomalies over the Niño3.4 region in the equatorial Pacific.

### C.2 Baselines and Ablations

We compare against the linear pattern scaling (LPS) baseline [Lütjens et al., 2025], which first regressing GMST on GHG, and then learning a point-wise linear mapping from GMST to the spatial temperature field. Despite its simplicity, LPS has been shown to outperform several deep-learning

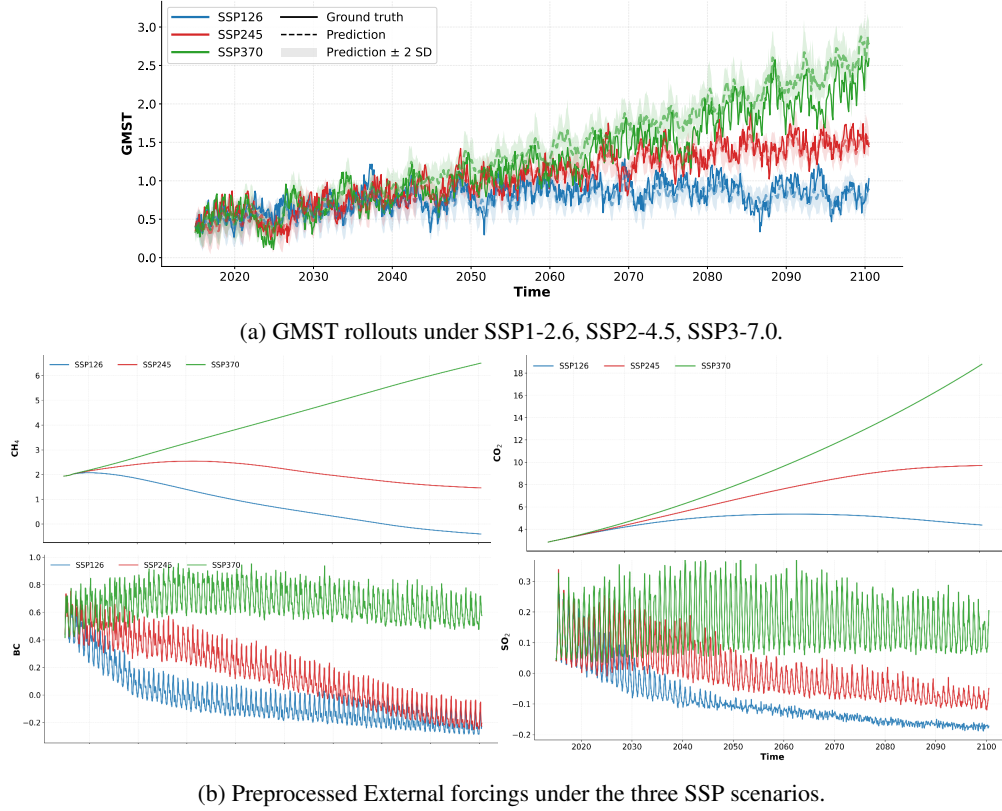


Figure 4: **GMST rollouts and forcings.** (a) GMST rollouts under SSP1-2.6 (blue), SSP2-4.5 (red), and SSP3-7.0 (green). Predicted GMST (dashed) is compared with NorESM2 (solid), with shading indicating  $\pm 2$  standard deviations across 50 trajectories. Our model is able to distinguish the scenario-dependent warming trends and captures the stronger warming under the held-out SSP3-7.0 scenario. (b)  $CO_2$ , spatially averaged  $CH_4$ ,  $BC$ , and spatially averaged  $SO_2$  forcings are shown for SSP1-2.6, SSP2-4.5, and SSP3-7.0 over the 85-year evaluation period. SSP1-2.6 and SSP2-4.5 represent lower- and intermediate-forcing pathways, respectively, while SSP3-7.0 represents a scenario with higher greenhouse gas and aerosol concentrations.

climate emulators. Our model better captures the mean GMST and Niño3.4 statistics, while LPS more accurately reproduces the GMST range. However, LPS substantially underestimates the Niño3.4 range and captures primarily the long-term trend. Consequently, it fails to reproduce the characteristic spectral variability of Niño3.4 time series, resulting in a near-zero PSD at interannual timescales (Fig. 2b) after detrending.

We also compare the full model with two ablations. *No Localized Aerosols* removes the localized aerosol assignment, allowing aerosol forcings from all locations to affect each local latent. *No  $\mathcal{L}_{z_g}$*  removes the alignment between the global latent and GMST. Without this alignment, its rollouts follow the particle selection procedure in Hickman et al. [2025]. Table 2 summarizes the rollout statistics, while Fig. 11 compares the temporal PSDs of GMST and Niño3.4. Removing the localized mapping improves the statistics for both indices, but suppresses temporal variability, producing an overly smooth Niño3.4 PSD and failing to capture its characteristic peak. Removing  $\mathcal{L}_{z_g}$  degrades the statistical scores and leads to larger long-term drift from the NorESM2 reference.

### C.3 Additional counterfactual experiments

We provide additional analyses of the forcing-perturbation experiments, including sampled temperature responses to regional aerosol perturbations (Fig. 12), single-step forcing-response pathways for each forcing variable (Fig. 13 – Fig. 16), and the GMST response to perturbations of the learned

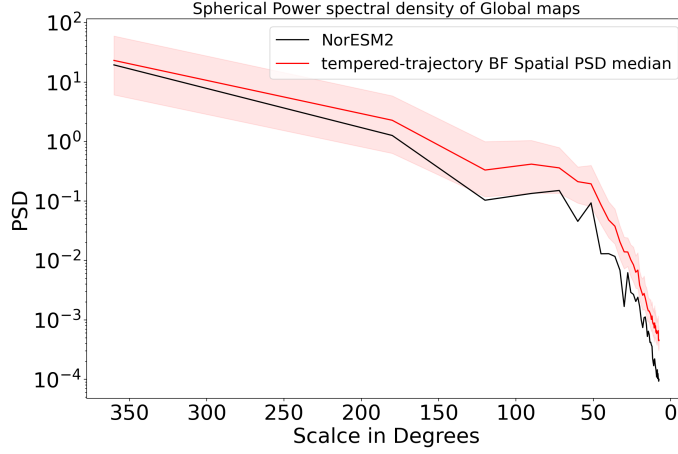


Figure 5: **Spherical power spectral density (PSD) of the time-averaged global surface temperature field.** For each trajectory, monthly surface temperature maps are first averaged over time, followed by a spherical harmonic decomposition of the resulting global HEALPix map. The spatial power spectrum is represented by the angular power spectrum  $C_\ell$ . The horizontal axis shows the approximate angular scale,  $360^\circ/(\ell + 1)$ , where smaller values correspond to finer spatial structures. The black curve denotes the NorESM2 reference, while the red curve shows the median PSD across model trajectories, with the shaded region indicating the 16th–84th percentile range. Both spectra are dominated by large-scale spatial structures, with power decreasing toward finer scales. The model closely reproduces the overall spectral slope and scale hierarchy of NorESM2, while it overestimates the power at fine spatial scales.

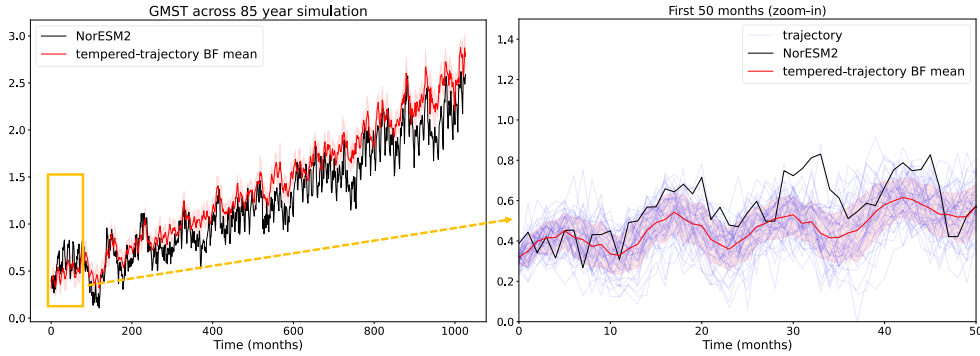


Figure 6: **GMST across 85 year rollouts under SSP3-7.0.** The left panel shows the global mean surface temperature (GMST) over the full 85-year autoregressive rollout. The right panel provides a detailed view of the first 50 months, showing the ensemble mean and standard deviation across 50 trajectories, together with all 50 individual sample trajectories.

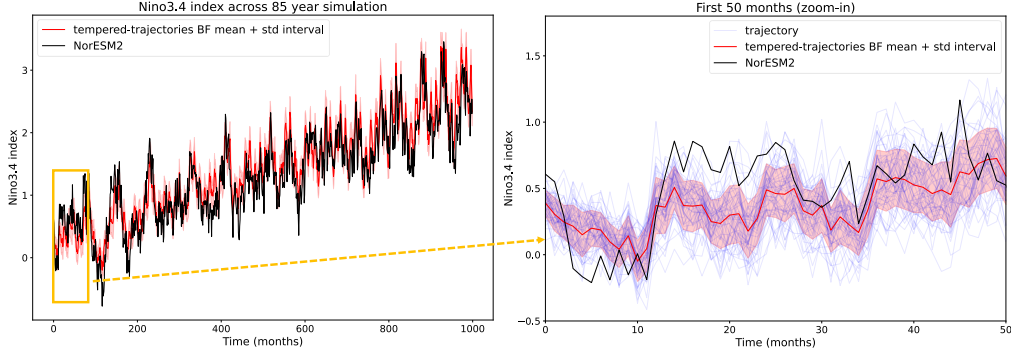


Figure 7: **Niño3.4 index across 85-year rollouts under SSP3-7.0.** The left panel shows the Niño3.4 index over the full 85-year autoregressive rollout. The right panel provides a detailed view of the first 50 months, showing the mean and standard deviation of 50 trajectories, and 50 individual sample trajectories.

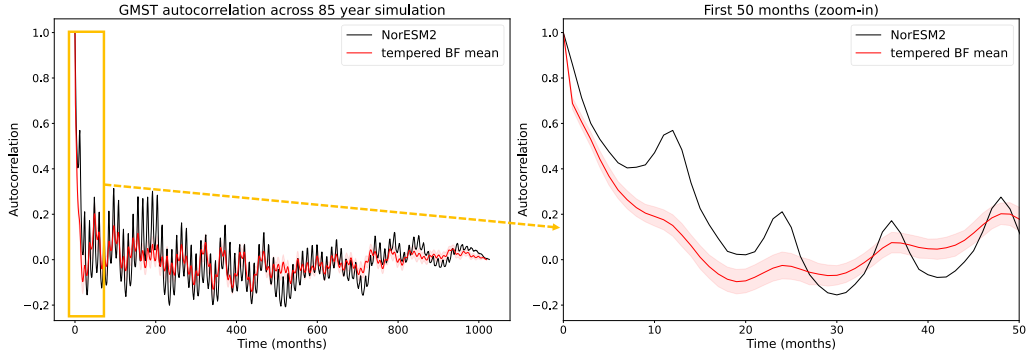


Figure 8: **GMST time-series autocorrelation under SSP3-7.0.** The left panel shows the autocorrelation of GMST over the full 85-year autoregressive rollout, comparing the model prediction with the NorESM2 ground truth. The right panel provides a detailed view of the first 50 months, highlighting the short-term temporal dependence and seasonal oscillations. Both time series are linearly detrended prior to computing the autocorrelation to remove the long-term warming trend and prevent it from dominating the autocorrelation at large time lags. The model captures the approximately annual oscillatory structure of the reference autocorrelation, while exhibiting weaker short-term persistence.

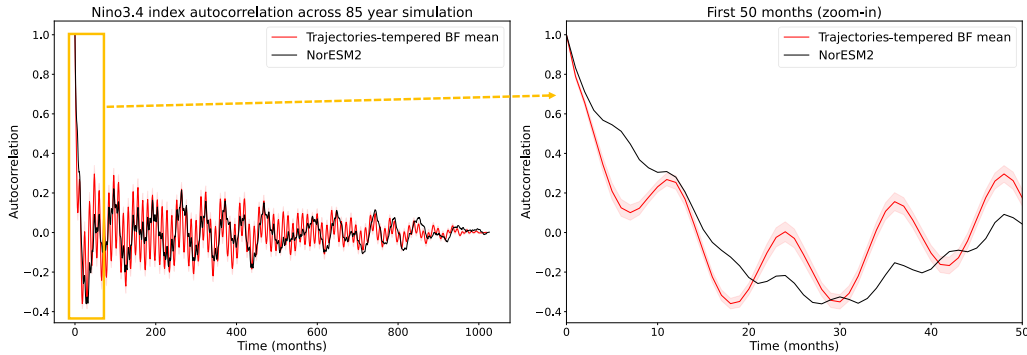


Figure 9: **Niño3.4 time-series autocorrelation under SSP3-7.0.** The left panel shows the autocorrelation of the Niño3.4 index over the full 85-year autoregressive rollout, comparing the model prediction with the NorESM2 ground truth. The right panel provides a detailed view of the first 50 months, highlighting the short-term temporal dependence and oscillatory variability. Both time series are linearly detrended prior to computing the autocorrelation to remove the long-term trend and prevent it from dominating the autocorrelation at large time lags. Compared with GMST, the model more closely reproduces the temporal dependence of the Niño3.4 index at one-year lag.

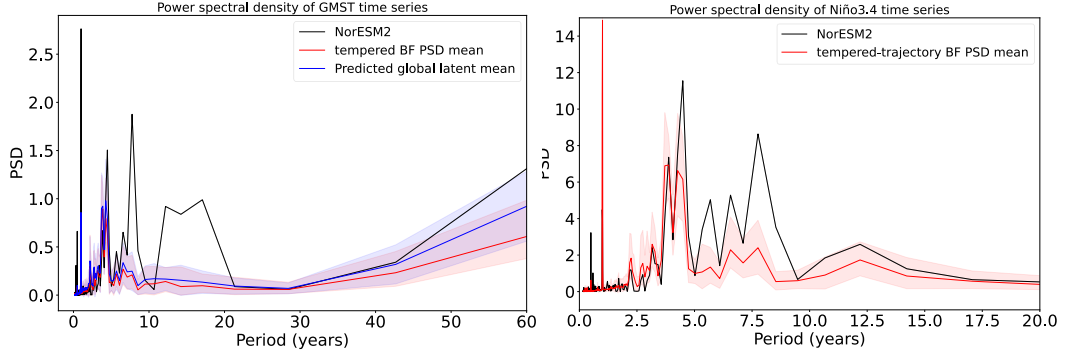


Figure 10: **Power spectral density (PSD) of the GMST and Niño3.4 index time series under SSP3-7.0.** The left panel shows the PSD of the GMST time series, while the right panel shows the PSD of the Niño3.4 index. The model PSD is computed across 50 autoregressive rollout trajectories, with the solid red line indicating the ensemble mean and the shaded region indicating  $\pm 1$  standard deviation. The black line denotes the NorESM2 ground-truth spectrum. In the GMST panel, the blue line additionally shows the PSD of the learned global climate latent, illustrating its temporal variability across different time scales. All time series are linearly detrended prior to PSD estimation. Global latent, compared with the spatial mean of the predicted temperature field, is better at capturing the long-term variability of GMST. The model reproduces the major spectral peaks of the Niño3.4 index, including the annual peak and variability at approximately 2 to 5 year periods.

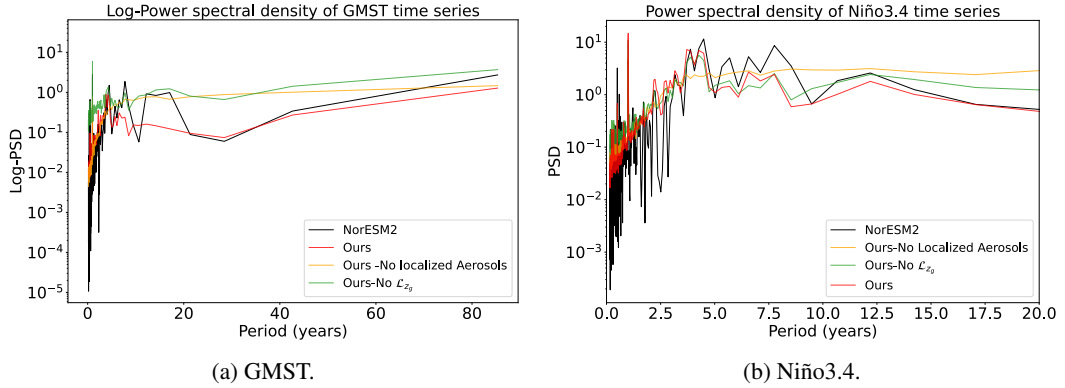


Figure 11: **Ablation study of the temporal power spectral density (PSD) under SSP3-7.0, shown on a logarithmic scale.** (a) GMST and (b) Niño3.4 PSDs from 85-year autoregressive rollouts. We compare the full model with variants without the local aerosol mapping and without the global-latent alignment loss  $\mathcal{L}_{zg}$ . Removing  $\mathcal{L}_{zg}$  amplifies temporal variability, especially over the low-frequency part. Removing the local mapping produces a smoother PSD and fails to capture the characteristic Niño3.4 peak.

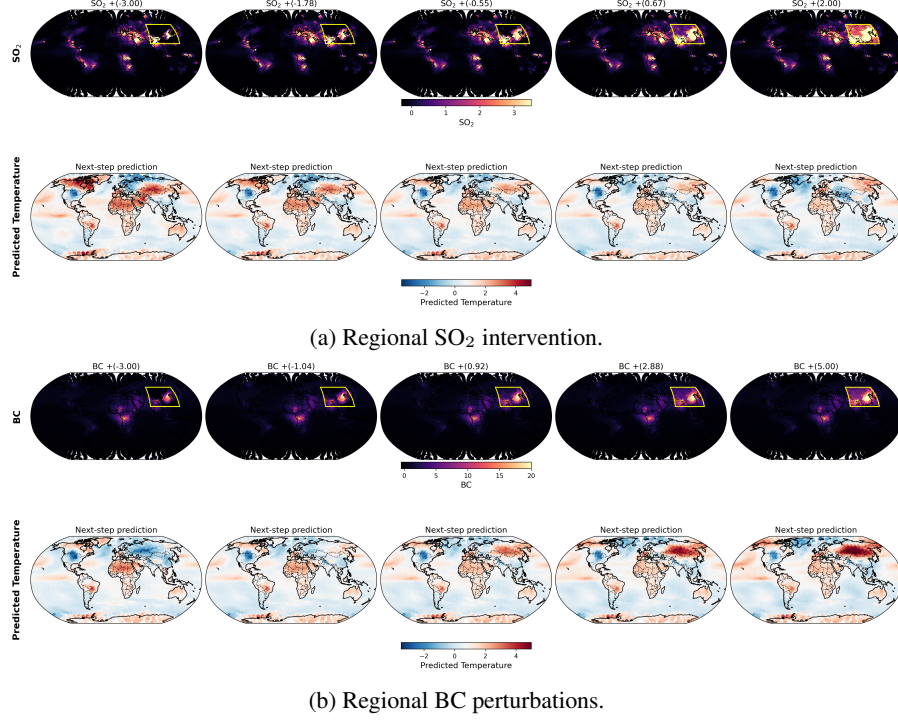


Figure 12: **Sample visualization of the temperature response to aerosol perturbations.** The upper row shows five different SO<sub>2</sub> perturbation levels applied within East Asia, and the lower row shows the corresponding next-step temperature predictions. Increasing SO<sub>2</sub> produces a cooling response, whereas increasing BC produces a warming response, with the strongest changes concentrated around the perturbed region.

global latent  $z_g$  (Fig. 18). Together, these analyses further characterize how global and regional forcing signals propagate through the learned representation to the predicted climate state.

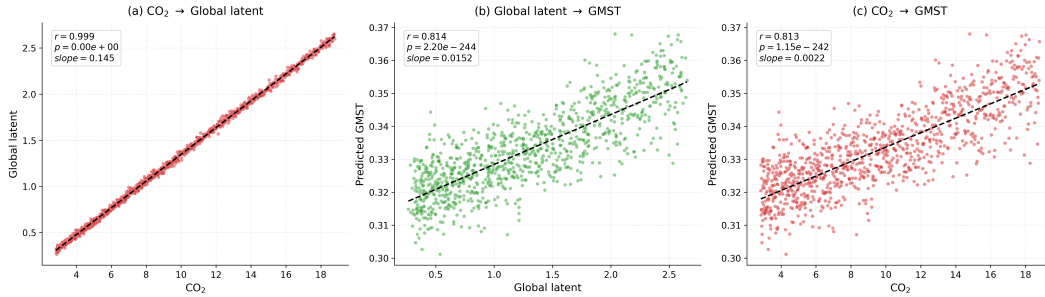


Figure 13: **Single-step response to CO<sub>2</sub> perturbations.** We evaluate the model under controlled CO<sub>2</sub> perturbations while keeping the input climate state and all other forcing variables fixed. Starting from a random selected initial condition, CO<sub>2</sub> is varied across its range of values observed under SSP3-7.0, and a single-step prediction is performed for each forcing value. (a) Relationship between the perturbed CO<sub>2</sub> concentration and the predicted global climate latent ( $\mu_{z_g}$ ). (b) Relationship between the predicted global latent and the corresponding predicted GMST. (c) Direct relationship between the perturbed CO<sub>2</sub> concentration and predicted GMST. Dashed lines indicate linear fits, with the Pearson correlation coefficient ( $r$ ),  $p$ -value, and fitted slope reported in each panel. The results show a nearly linear response of the global latent to CO<sub>2</sub> perturbations, which is subsequently reflected in the predicted GMST.

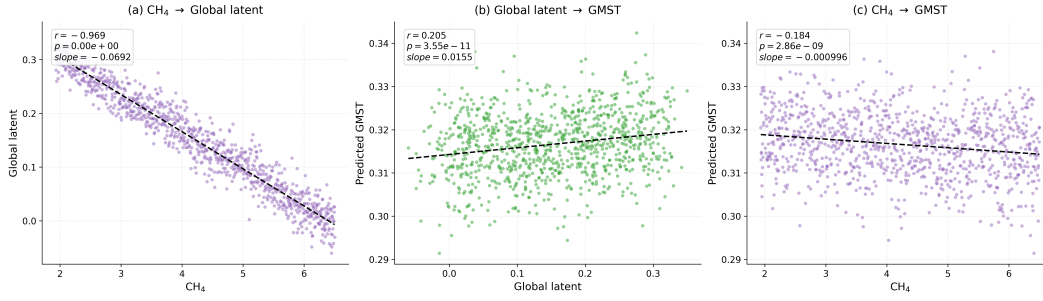


Figure 14: **Single-step response to  $\text{CH}_4$  perturbations.** Compared with  $\text{CO}_2$ ,  $\text{CH}_4$  exhibits a weaker modulation of the learned global latent, and this variation is propagated even more weakly to the next-step GMST prediction, resulting in a substantially weaker direct  $\text{CH}_4$ –GMST response. The cooling response to increased  $\text{CH}_4$  reflects a limitation of the current model in disentangling the effects of strongly correlated global forcings and remains an area for further improvement.

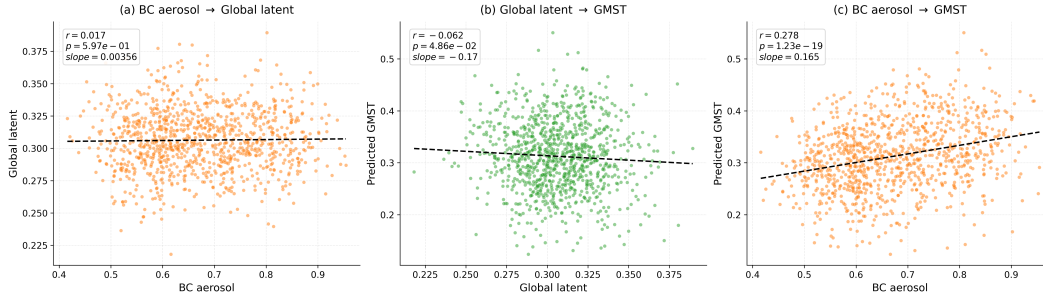


Figure 15: **Single-step response to BC perturbations.** In contrast to the global greenhouse-gas forcings, BC shows almost no modulation of the learned global latent ( $r = 0.017$ ), while exhibiting a clearer positive response in the predicted GMST ( $r = 0.278$ ). This suggests that the modeled BC temperature response is primarily mediated through the local climate latents rather than the global latent.

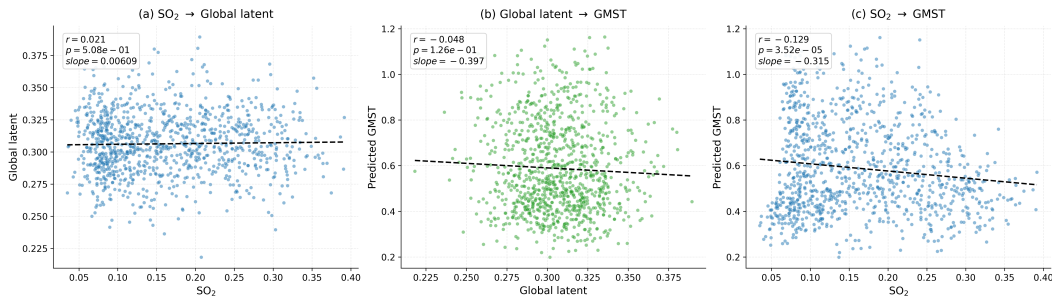


Figure 16: **Single-step response to  $\text{SO}_2$  perturbations.** Similar to BC,  $\text{SO}_2$  shows almost no modulation of the learned global latent ( $r = 0.021$ ), while exhibiting a weak negative response in the predicted GMST ( $r = -0.129$ ). The global latent responds to global forcing perturbations but remains insensitive to aerosol perturbations, providing evidence that the model separates global and localized forcings through distinct causal pathways.

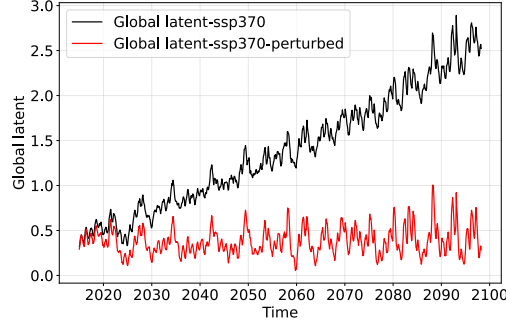


Figure 17: **Global latent response to forcing perturbations in long-term rollout.** Global latent trajectories under SSP3-7.0 with original forcings (black) and with  $\text{CO}_2$  and  $\text{CH}_4$  fixed at their 2015 levels (red), showing a substantially suppressed long-term increase under the perturbation.

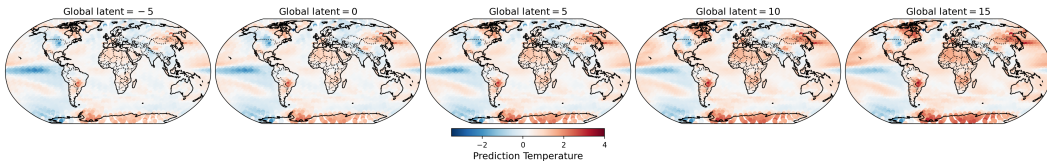


Figure 18: **Temperature response to interventions on the predicted global climate latent.** We directly intervene on the predicted global latent mean ( $\mu_{z_g}$ ) by setting it to  $-5$ ,  $0$ ,  $5$ ,  $10$ , and  $15$ , while keeping the input climate history and external forcing conditions unchanged. For each intervention, the prescribed global latent is passed to the local climate transition model, and the resulting local latent state is decoded into the predicted surface temperature field. Each panel shows the prediction for the same input sample under a different value of the global latent. As the global latent increases, the predicted temperature field exhibits a progressively stronger warming response over much of the globe. This sensitivity experiment indicates that the learned global latent modulates the large-scale temperature response.

Table 1: Model and training hyperparameters used in the experiments. We denote an MLP with hidden dimension  $a$  and  $b$  layers as  $\text{MLP}(a, b)$ , and use  $\text{Linear}(a, b)$  analogously for the corresponding linear layer. The model is trained with RMSProp (lr =  $10^{-4}$ , batch size 128) for up to  $10^5$  iterations. The auxiliary loss term incorporates the continuous ranked probability score (CRPS) and a spatio-temporal spectral loss on the predicted temperature.

| Parameter                                   | Value                               |
|---------------------------------------------|-------------------------------------|
| <i>Latent representation</i>                |                                     |
| Number of spatial grid cells $d_x$          | 3072                                |
| Number of local latent variables $d_z$      | 48                                  |
| Number of global latent variables $d_{z_g}$ | 1                                   |
| Temporal context length $\tau$              | 5 months                            |
| <i>Encoder and decoder</i>                  |                                     |
| Encoder                                     | MLP(8,2)                            |
| Decoder                                     | MLP(8,2)                            |
| Activation function                         | LeakyReLU                           |
| Encoder distribution                        | Gaussian                            |
| Decoder distribution                        | Gaussian                            |
| <i>Latent dynamics</i>                      |                                     |
| Local transition dynamics                   | MLP(8,2)                            |
| Global transition dynamics                  | $\text{Linear}(d_z \times \tau, 1)$ |
| Transition distribution                     | Gaussian                            |
| Global forcings                             | CO <sub>2</sub> , CH <sub>4</sub>   |
| Spatial forcings                            | BC, SO <sub>2</sub>                 |
| <i>Loss coefficients</i>                    |                                     |
| KL coefficient $\lambda_{\text{KL}}$        | 1                                   |
| GMST coefficient $\lambda_{\text{GMST}}$    | $10^3$                              |
| CRPS coefficient $\lambda_{\text{CRPS}}$    | 1                                   |
| Spatial spectral coefficient                | $10^3$                              |
| Temporal spectral coefficient               | $10^3$                              |
| <i>Optimization</i>                         |                                     |
| Optimizer                                   | RMSProp                             |
| Initial learning rate                       | $10^{-4}$                           |
| Batch size                                  | 128                                 |
| Maximum training iterations                 | $10^5$                              |
| Validation interval                         | 100 iterations                      |
| Early-stopping patience                     | 5000 iterations                     |
| <i>Single-parent constraint (ALM)</i>       |                                     |
| Initial penalty $\mu_0$                     | $10^4$                              |
| Penalty multiplier                          | 1.2                                 |
| $\omega_\gamma$                             | 0.01                                |
| $\omega_\mu$                                | 0.9                                 |
| Constraint threshold                        | $10^{-2}$                           |
| Minimum convergence interval                | 1000                                |
| <i>Transition sparsity constraint (ALM)</i> |                                     |
| Initial penalty $\mu_0$                     | $10^{-2}$                           |
| Penalty multiplier                          | 1.2                                 |
| $\omega_\gamma$                             | 0.01                                |
| $\omega_\mu$                                | 0.95                                |
| Constraint threshold                        | $10^{-4}$                           |
| Minimum convergence interval                | 1000                                |
| Graph binarization threshold                | 0.5                                 |
| <i>Rollout</i>                              |                                     |
| Number of particles                         | 50                                  |
| Candidate samples per particle              | 10                                  |

Table 2: **Statistics of long-term climate rollouts under SSP3-7.0 forcing.** Starting from the climate state on 2015-01-01, each model is rolled out autoregressively for 85 years, conditioned at each time step on the corresponding SSP3-7.0 external forcings. The particle selection follows 2.2. The left section shows the mean, standard deviation, range of the GMST, and the same for the Niño3.4 index on the right. The first row shows the reference statistics computed from the ground-truth NorESM2 SSP3-7.0 simulation, followed by comparison method, and then ablations. For all metrics, being closer to the ground truth is better.

|                                              | GMST          |               |               | Niño3.4       |               |               |
|----------------------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|
|                                              | Mean          | Std. Dev.     | Range         | Mean          | Std. Dev.     | Range         |
| Ground truth                                 | 1.2403        | 0.5771        | 2.5144        | 1.3662        | 0.7751        | 4.2305        |
| Linear Pattern Scaling Lütjens et al. [2025] | 1.4711        | 0.7009        | <b>2.4247</b> | 1.5677        | <b>0.7383</b> | 2.5535        |
| Ours-No Localized Aerosols                   | <b>1.3281</b> | <b>0.5327</b> | 2.7091        | <b>1.4013</b> | 0.6702        | <b>4.3710</b> |
| Ours-No $\mathcal{L}_{z_g}$                  | 1.0015        | 0.6247        | 4.8010        | 1.1235        | 0.6663        | 4.7029        |
| Ours                                         | 1.4107        | 0.6573        | 3.2710        | 1.4692        | 0.8205        | 4.7190        |