
In-distribution skill does not diagnose transfer: a controlled ocean testbed for scientific representations

Huan Zhang

LMD/IPSL, CNRS, ENS, Université PSL,
École Polytechnique, Institut Polytechnique de Paris, Sorbonne Université,
Paris, France
huan.zhang@lmd.ipsl.fr

Abstract

Scientific representations are commonly selected by in-distribution predictive accuracy, but this need not determine transfer after physical regime change. We compare four data-driven predictors of the same five-channel North Atlantic ocean state: Euclidean and spatially regularized regression, low-rank EOF/POP dynamics, and a nonlinear latent dynamical emulator. All are trained on preindustrial control data and evaluated at 1-, 5-, and 10-year leads under historical and SSP3-7.0 forcing. Their in-distribution and out-of-distribution rankings are reversed: the nonlinear emulator is modestly worse on preindustrial test data at long leads but generalizes best after a late-century reorganization marked by near-suppression of Labrador Sea deep convection. POP is intermediate, while direct regressions fail most strongly. Attribution analysis suggests that state-evolution objectives preserve multivariate physical relationships weakly constrained by the scalar prediction task but relevant after regime change. These results motivate evaluating scientific representations across physically meaningful dynamical shifts, not only held-out samples from the training regime. These results suggest when both regimes cannot be sampled, predictors should preserve generalizable physical mechanisms and extrapolate along a physically plausible manifold.

1 Introduction

Data-driven weather and climate emulators are increasingly used not only for weather prediction but also for multidecadal climate simulation under greenhouse gas forcing, where they must operate in a climate that differs from those represented during training [Dheeshjith et al., 2025, Watt-Meyer et al., 2025, Kochkov et al., 2024], therefore an out-of-distribution (OOD) problem. Recent studies show that models with strong in-distribution skill can develop systematic biases under warming or underestimate forced responses [Rasp et al., 2018, Beucler et al., 2024, Zhang and Merlis, 2025, Dheeshjith et al., 2025]. Recent approaches have sought to improve generalization through physically informed feature transformations or model design [Beucler et al., 2024, Stanley-Clamp et al., 2026], explicit conditioning on external forcing [Watt-Meyer et al., 2025], or training on simulations spanning a broader range of climate states or forcing levels [Clark et al., 2022, 2025]. More broadly, growing evidence in machine learning indicates that off-support is shaped not only by data coverage but also by architecture and representation: networks with similar interpolation error can extrapolate differently because their parameterization, activation structure, or learned basis privileges different directions of variation [Xu et al., 2021, Tolmachev and Engel, 2025, Xie and Luo, 2026, Shikhman, 2026]. We therefore ask: *which representational constraints are invisible to in-distribution predictive skill but become consequential under physical regime change?*

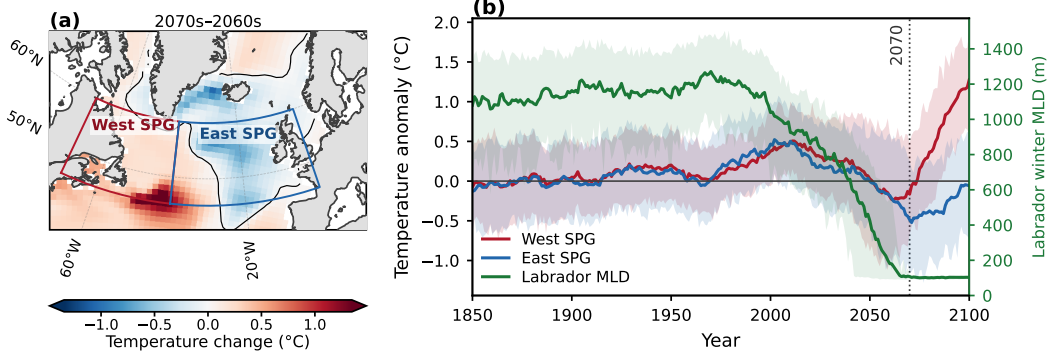


Figure 1: CESM2 SPG reorganization. (a) Late-century 0–200 m temperature change showing the west–east dipole. (b) Regional temperature and Labrador winter mixed-layer depth.

The North Atlantic subpolar gyre (SPG) provides such a test because its coupled temperature, salinity, deep-convection, and circulation dynamics can undergo substantial reorganization under warming [Sgubin et al., 2017, Swingedouw et al., 2021, Gu et al., 2024]. In the CESM2 Large Ensemble used here [Danabasoglu et al., 2020, Rodgers et al., 2021], Labrador Sea winter convection becomes nearly suppressed late in the century, reorganizing coupled temperature, salinity, and circulation relationships. Using identical five-channel states and PI training data, we compare direct regularized predictors, EOF/POP state dynamics [Hasselmann, 1988, von Storch et al., 1995], and a nonlinear latent-state emulator. Despite similar PI skill, their transfer ranking reverses after this reorganization; finite-change attribution diagnoses the pathways associated with the divergence.

2 Testbed

We use a 2000-year CESM2 pre-industrial (PI) control [Danabasoglu et al., 2020] for training and in-distribution evaluation, and the 100-member CESM2 Large Ensemble under historical and SSP3-7.0 forcing [Rodgers et al., 2021] for OOD evaluation. Inputs are February–March–April (FMA)-mean North Atlantic–Arctic states north of 40°N with five channels:

$$T_{0-200}, T_{200-1000}, S_{0-200}, S_{200-1000}, \text{SSH},$$

and the target is area-weighted SPG 0–200 m temperature at 1-, 5-, and 10-year leads. PI years 1–1400/1401–1700/1701–2000 are used for training/validation/test; hyperparameters use validation only. We then **freeze the PI-trained predictors and transfer them without adaptation** to 1850–2100. Full data protocols, preprocessing, and evaluation definitions are given in Appendix A. CESM2 North Atlantic state undergoes a marked reorganization around 2070. Labrador Sea winter mixed layers become extremely shallow, falling to approximately 9.4% of their PI mean, consistent with near-suppression of deep convection [Marshall and Schott, 1999]. This is accompanied by a pronounced west–east temperature dipole (Fig. 1): the western SPG recovers while the eastern SPG remains cold.

3 The representation ladder: target-specific prediction versus state evolution

We compare predictors with the same five-channel input but progressively stronger constraints on what must be represented before the scalar target is predicted: direct scalar regression, low-rank linear state evolution, and learned nonlinear state evolution. Implementation details are in Appendix B.

Direct prediction with explicit priors. The regularized linear predictors take the form

$$\hat{m} = \arg \min_m \left\{ \frac{1}{nd} \|Y - Xm\|_2^2 + \alpha \|m\|^2 + \gamma \|Dm\|_2^2 \right\}, \quad \hat{y} = \hat{m}^\top x \quad (1)$$

where D is the discrete spatial-gradient operator. We define L^2 model when $\gamma = 0$ corresponding to isotropic shrinkage of the coefficient field, and H^1 model when $\gamma > 0$, additionally penalizing spatial roughness.

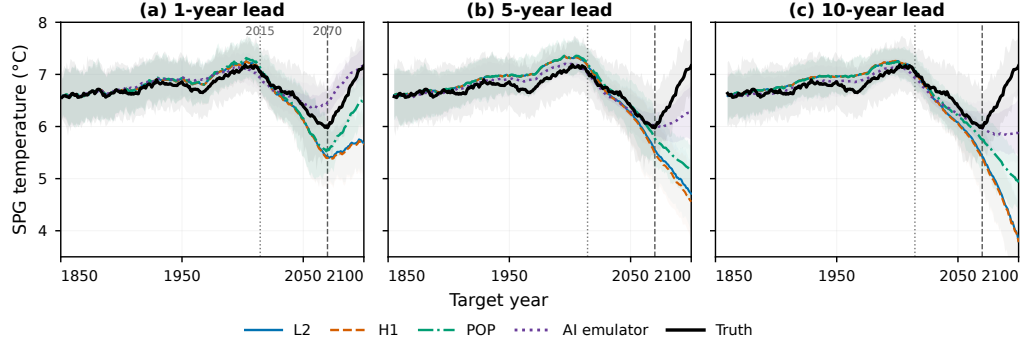


Figure 2: PI-trained predictions under historical and SSP3-7.0 forcing at 1-, 5-, and 10-year leads. Curves show ensemble means and gray shading the truth spread; vertical lines mark SSP3-7.0 onset (2015) and the late-century reorganization (2070).

Low-rank dynamical state representation POP instead first projects the state onto EOFs and fits its Lead- L evolution matrix A_L and read-out vector c_L :

$$z_t = E^\top x_t, \quad \hat{z}_{t+L} = A_L z_t, \quad \hat{y}_{t+L} = c_L^\top \hat{z}_{t+L}, \quad (2)$$

constraining a low-rank multivariate state evolution rather than directly optimizing the scalar target.

Learned nonlinear latent state The nonlinear emulator replaces the fixed EOF basis with a learned 64-dimensional autoencoder representation and recursively evolves it,

$$x_t \xrightarrow{E_\phi} z_t \xrightarrow{F_\theta^L} z_{t+L} \xrightarrow{D_\phi} \hat{x}_{t+L} \rightarrow \hat{y}_{t+L} \quad (3)$$

The autoencoder is trained to reconstruct the five-channel state and frozen before transition training. Predictions are averaged across ten independently trained transition models.

4 Results

4.1 In-distribution skill does not predict transfer across the regime shift

Table 1 shows RMSE and member-averaged late-SSP RMSE; evaluation details are in Appendix A and C. On the PI-control test data, L^2/H^1 achieve the lowest RMSEs, with POP only slightly worse and the nonlinear emulator worst, especially at longer leads for a cold bias (not shown). This ordering reverses in 2070–2100 at every lead:

Table 1: Prediction RMSE on PI-control test data and the late forced period (2070–2100). Lower is better.

Model	Lead 1		Lead 5		Lead 10	
	PI	Late	PI	Late	PI	Late
L2	.214	1.072	.261	1.607	.294	2.098
H1	.214	1.103	.261	1.704	.294	2.132
POP	.217	.615	.270	1.297	.301	1.429
Emulator	.228	.466	.300	.645	.339	.873

$$\text{AI-emulator} < \text{POP} < L^2/H^1 \quad (\text{RMSE}).$$

Fig. 2 shows that the separation is not a uniform consequence of warming: the PI-trained models remain similar through much of the historical and early-SSP period, then diverge sharply near the late-century reorganization. All develop a cold bias, but the direct regressions continue cooling most strongly, POP captures part of the recovery, and the emulator follows the rebound most closely. Thus, nearly indistinguishable PI predictive skill does not determine transfer once the dynamically occupied regime changes.

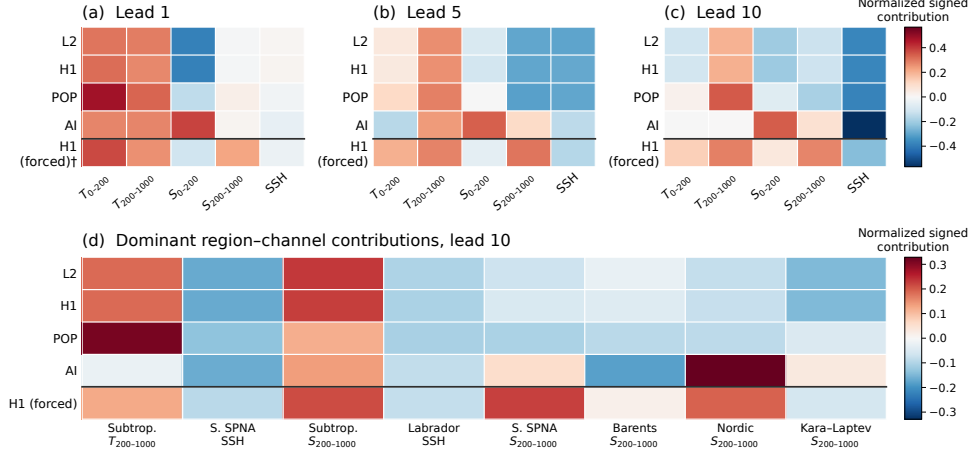


Figure 3: Normalized contributions to the 2070–2100 response. (a–c) Channel contributions at 1-, 5-, and 10-year leads; POP emphasizes subsurface temperature, while the emulator emphasizes salinity. (d) Dominant lead-10 region–channel contributions. † [H^1 (forced)] is a diagnostic reference.]

4.2 Representation-dependent pathways under late-climate transfer

Fig. 3 examines which components of the late-century state change contribute to each model’s predicted response. L^2 and H^1 remain nearly identical, whereas POP and the emulator rely on distinct multivariate pathways. POP places a stronger positive contribution on $T_{200-1000}$, particularly in the subtropical/southern North Atlantic, while the emulator places greater weight on salinity. At lead 10, its clearest distinction is a positive contribution from Nordic-Seas subsurface salinity, opposite in sign to the PI-trained direct regressions and closer to the forced-trained H^1 diagnostic. These differences suggest that models trained to evolve the multivariate state retain relationships that are weakly constrained by the PI scalar prediction task but become relevant after reorganization. The attributions are predictive rather than causal. See Appendix D and E for details.

5 Discussion and limitations

The attribution suggests that POP and the nonlinear emulator transfer through different physical pathways. POP’s improvement is associated with subsurface temperature, including the eastern SPG/North Atlantic Current gateway (more obvious in lead 5, not shown) and the subtropical–southern North Atlantic. A plausible interpretation is that its low-rank linear state dynamics preserve approximately advective heat transport and vertical coupling between T_{0-200} and $T_{200-1000}$, including entrainment. The emulator instead relies more strongly on salinity around the high-latitude gateways; at lead 10, its clearest distinction is a positive Nordic-Seas subsurface-salinity contribution, opposite to the PI-trained direct regressions and closer to the forced-data H^1 diagnostic. This is consistent with salinity–convection–circulation coupling, although the attribution is predictive rather than causal. Thus the two state-evolution models appear to retain complementary multivariate relationships that direct scalar regression can discard. As a control, expanded forced-climate support with scalar CO_2 encoding improves some pre-transition forecasts but does not rescue late-regime skill (Appendix F), showing that scenario information alone is insufficient to resolve the post-reorganization failure.

More broadly, the experiment suggests that in-distribution accuracy underdetermines how a predictor is continued beyond the occupied training regime. Full-state objectives constrain this continuation: POP through a low-rank linear evolution and the emulator through a learned nonlinear latent evolution. Neither constraint is universally preferable—POP emphasizes approximately linear transport processes, while the emulator is more flexible but exhibits substantial seed dependence—but both outperform direct scalar regression after the circulation reorganizes. This motivates designing architectures, objectives, and rollout procedures that constrain physically plausible state evolution, rather than relying only on target-specific statistical correlations.

Several limitations temper this interpretation. Even the best model remains substantially biased in the late climate, so the result is a difference in degree of failure rather than successful extrapolation. We study one climate model and one forcing scenario, and do not isolate representation from architecture or training objective. The inferred attributions are predictive sensitivities, not causal response functions, so the observed ordering should not be interpreted as universal.

Acknowledgments and Disclosure of Funding

This paper is ClimTip contribution No. X. The ClimTip project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No. 101137601. This study benefited from the computational resources of the IPSL Data and Computing Center ESPRI, including the Spirit computing cluster. ESPRI is supported by CNRS, Sorbonne Université, CNES, and École Polytechnique. The authors also acknowledge the high-performance computing resources provided by GENCI-IDRIS (Grant No. 2025-AD010116382).

References

- Tom Beucler, Pierre Gentine, Janni Yuval, Ankitesh Gupta, Liran Peng, Jerry Lin, Sungduk Yu, Stephan Rasp, Fiaz Ahmed, Paul A. O’Gorman, J. David Neelin, Nicholas J. Lutsko, and Michael Pritchard. Climate-invariant machine learning. *Science Advances*, 10(6):eadj7250, 2024. doi: 10.1126/sciadv.adj7250.
- Spencer K. Clark, Noah D. Brenowitz, Brian Henn, Anna Kwa, Jeremy McGibbon, W. Andre Perkins, Oliver Watt-Meyer, Christopher S. Bretherton, and Lucas M. Harris. Correcting a 200 km resolution climate model in multiple climates by machine learning from 25 km resolution simulations. *Journal of Advances in Modeling Earth Systems*, 14(9):e2022MS003219, 2022. doi: <https://doi.org/10.1029/2022MS003219>. URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003219>. e2022MS003219 2022MS003219.
- Spencer K. Clark, Oliver Watt-Meyer, Anna Kwa, Jeremy McGibbon, Brian Henn, W. Andre Perkins, Elynn Wu, Lucas M. Harris, and Christopher S. Bretherton. Ace2-som: Coupling an ml atmospheric emulator to a slab ocean and learning the sensitivity of climate to changed co2. *Journal of Geophysical Research: Machine Learning and Computation*, 2(4):e2024JH000575, 2025. doi: <https://doi.org/10.1029/2024JH000575>. URL <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024JH000575>. e2024JH000575 2024JH000575.
- Gokhan Danabasoglu, Jean-François Lamarque, Julio Bacmeister, David A. Bailey, Alice K. DuVivier, Jim Edwards, Louisa K. Emmons, John Fasullo, Rolando Garcia, Andrew Gettelman, et al. The community earth system model version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001916, 2020. doi: 10.1029/2019MS001916.
- Surya Dheeshjith, Adam Subel, Alistair Adcroft, Julius Busecke, Carlos Fernandez-Granda, Shubham Gupta, and Laure Zanna. Samudra: An AI global ocean emulator for climate. *Geophysical Research Letters*, 52(10):e2024GL114318, 2025. doi: 10.1029/2024GL114318.
- Qinxue Gu, Melissa Gervais, Gokhan Danabasoglu, Who M. Kim, Frederic Castruccio, Elizabeth Maroon, and Shang-Ping Xie. Wide range of possible trajectories of north atlantic climate in a warming world. *Nature Communications*, 15:4221, 2024. doi: 10.1038/s41467-024-48401-2.
- Klaus Hasselmann. PIPs and POPs: The reduction of complex dynamical systems using principal interaction and oscillation patterns. *Journal of Geophysical Research: Atmospheres*, 93(D9): 11015–11021, 1988. doi: 10.1029/JD093iD09p11015.
- Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, and Matthew Willson. Neural general circulation models for weather and climate. *Nature*, 632:1060–1066, 2024. doi: 10.1038/s41586-024-07744-y.
- John Marshall and Friedrich Schott. Open-ocean convection: Observations, theory, and models. *Reviews of Geophysics*, 37(1):1–64, 1999. doi: 10.1029/98RG02739.

- Stephan Rasp, Michael S. Pritchard, and Pierre Gentine. Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39):9684–9689, 2018. doi: 10.1073/pnas.1810286115.
- Keith B. Rodgers, Sun-Seon Lee, Nan Rosenbloom, Axel Timmermann, Gokhan Danabasoglu, Clara Deser, Jim Edwards, J.-E. Kim, Isla R. Simpson, K. Stein, Malte F. Stuecker, R. Yamaguchi, T. Bódai, E.-S. Chung, L. Huang, W. M. Kim, J.-F. Lamarque, D. L. Lombardozzi, W. R. Wieder, and Stephen G. Yeager. Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, 12:1393–1411, 2021. doi: 10.5194/esd-12-1393-2021.
- Giovanni Sgubin, Didier Swingedouw, Sybren Drijfhout, Yves Mary, and Amine Bennabi. Abrupt cooling over the north atlantic in modern climate models. *Nature Communications*, 8:14375, 2017. doi: 10.1038/ncomms14375.
- Lennon Shikhman. Diagnosing failure modes of neural operators across diverse pde families, 2026. URL <https://arxiv.org/abs/2601.11428>.
- Bradley Stanley-Clamp, Anson Lei, Hannah M. Christensen, and Ingmar Posner. No epoch like the present: Robust climate emulation requires out-of-distribution generalisation. *arXiv preprint arXiv:2605.22248*, 2026.
- Didier Swingedouw, A. Bily, C. Esquerdo, L. F. Borchert, G. Sgubin, J. Mignot, and M. Menary. On the risk of abrupt changes in the north atlantic subpolar gyre in CMIP6 models. *Annals of the New York Academy of Sciences*, 1504(1):187–201, 2021. doi: 10.1111/nyas.14659.
- Pavel Tolmachev and Tatiana A Engel. Single-unit activations confer inductive biases for emergent circuit solutions to cognitive tasks. *Nature Machine Intelligence*, 7(10):1742–1754, 2025.
- Hans von Storch, Gerd Bürger, Reiner Schnur, and Jin-Song von Storch. Principal oscillation patterns: A review. *Journal of Climate*, 8(3):377–400, 1995. doi: 10.1175/1520-0442(1995)008<0377:POPAR>2.0.CO;2.
- Oliver Watt-Meyer, Brian Henn, Jeremy McGibbon, Spencer K. Clark, Anna Kwa, W. Andre Perkins, Elynn Wu, Lucas Harris, and Christopher S. Bretherton. ACE2: Accurately learning subseasonal to decadal atmospheric variability and forced responses. *npj Climate and Atmospheric Science*, 8: 205, 2025. doi: 10.1038/s41612-025-01090-0.
- Runlong Xie and An Luo. Frequency bias and ood generalization in neural operators under a variable-coefficient wave equation, 2026. URL <https://arxiv.org/abs/2605.12997>.
- Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S. Du, Ken ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks, 2021. URL <https://arxiv.org/abs/2009.11848>.
- Bosong Zhang and Timothy M. Merlis. The equilibrium response of atmospheric machine-learning models to uniform sea surface temperature warming. *arXiv preprint arXiv:2510.02415*, 2025.

A Data, preprocessing, and evaluation protocol

This appendix records the exact state construction and evaluation conventions used by the representation-transfer experiments. The purpose is to make clear which information is shared across the four predictors and which choices belong to a particular representation.

A.1 CESM2 simulations and seasonal sampling

We use CESM2 pre-industrial control (PI) and Large Ensemble simulations [Danabasoglu et al., 2020, Rodgers et al., 2021]. The PI record contains 2000 annual samples. The forced evaluation uses 100 CESM2 Large Ensemble members spanning historical forcing and SSP3–7.0 from 1850 through 2100. Each state is an annual February–March–April (FMA) ocean mean. All forecast windows are indexed by *target year*; a target at year t and lead L uses the source state from year $t - L$.

The primary PI split is chronological:

$$\mathcal{T}_{\text{train}} = 1:1400, \quad \mathcal{T}_{\text{val}} = 1401:1700, \quad \mathcal{T}_{\text{test}} = 1701:2000,$$

with lead-safe source/target pairs formed separately for $L \in \{1, 5, 10\}$ yr. Predictive regularization and model selection use PI validation only; the POP EOF rank is chosen separately by representation-capacity (Appendix C). The frozen PI-trained models are applied without adaptation to all 100 forced members. According to the ensemble-mean SPG temperature (black line in Figure 1), we define the target-year epochs

historical : 1850:2014, early SSP : 2015:2050, transition : 2051:2069, late SSP : 2070:2100.

A.2 Five-channel North Atlantic state

The gridded predictor covers ocean points north of 40°N on a 99×87 regrided ocean grid and contains five physical channels,

$$T_{0-200\text{m}}, \quad T_{200-1000\text{m}}, \quad S_{0-200\text{m}}, \quad S_{200-1000\text{m}}, \quad \text{SSH}.$$

Temperature and salinity are vertically averaged over the indicated depth ranges. Channel-specific valid-ocean masks are retained rather than intersected into a single common subsurface mask. Concatenating all valid features gives $d = 9,131$ predictor coordinates. In this appendix we denote the raw-physics input variable and target variable with channel units as $x_i(t)$ and $y_i(t)$.

A.3 Rotated-pole regridding and latitude weighting

All five ocean fields are represented on the same rotated and regrided North Atlantic–Arctic grid before model fitting. The native CESM2 fields are first expressed in a rotated-pole coordinate system and then regrided to the 99×87 grid used throughout this study. We denote the coordinates of this numerical grid by (rlat, rlon). The rotation places the analysis domain away from the geographic pole, giving a substantially more regular grid over the North Atlantic–Arctic domain.

It is important to distinguish the rotated-grid latitude rlat from geographic latitude ϕ . Whenever a cosine weight is used to represent grid-cell area or the corresponding grid metric in our numerical pipeline, we use

$$w_{ij} = \cos(\text{rlat}_{ij}), \quad (4)$$

rather than $\cos(\phi_{ij})$. Thus “latitude weighting” below refers to the latitude coordinate of the rotated and regrided grid unless explicitly stated otherwise. This convention is used consistently for area-weighted scalar quantities and for spatial weighting operators defined on the numerical grid.

Geographic latitude and longitude are retained separately and are used when defining physically meaningful regions, such as the SPG, Labrador Sea, Nordic Seas, and the regional partition used in the attribution analysis. They should therefore not be confused with (rlat, rlon), which specify the geometry on which the models and numerical spatial operators are implemented.

A.4 SPG prediction target

Using the rotated-grid area weights defined in Appendix A.3, the scalar target is the FMA 0–200 m temperature over the SPG box $45^\circ\text{--}60^\circ\text{N}$, $70^\circ\text{--}20^\circ\text{W}$,

$$y(t) = \frac{\sum_{(i,j) \in \Omega_{\text{SPG}}} T_{0-200\text{m}}^{\text{FMA}}(t, i, j) w_{ij}}{\sum_{(i,j) \in \Omega_{\text{SPG}}} w_{ij}}. \quad (5)$$

A.5 Definition of RMSE in Table 1

For CESM2-Large ensemble model predictions (with units $^\circ\text{C}$) $\hat{y}_{m,t}$ and truth $y_{m,t}$, first calculate an RMSE over time t for every climate member with index m :

$$\text{RMSE}_m = \sqrt{\frac{1}{N_T} \sum_{t \in T} (\hat{y}_{m,t} - y_{m,t})^2}.$$

Then average those RMSEs across the 100 members:

$$\text{RMSE}_{\text{table}} = \frac{1}{100} \sum_{m=1}^{100} \text{RMSE}_m$$

For the AI emulator, the ten network predictions are first averaged separately for every climate member:

$$\hat{y}_{m,t}^{AI} = \frac{1}{10} \sum_{s=1}^{10} \hat{y}_{s,m,t},$$

and then the same memberwise-RMSE formula is applied.

For PI-control, the RMSE is simply

$$\text{RMSE}_{PI} = \sqrt{\frac{1}{N} \sum_i (\hat{y}_i - y_i)^2}.$$

A.6 Complete prediction results

In Table 1 we predict the ten-emulator ensemble mean. To quantify uncertainty associated with the finite set of neural transition models, we bootstrap the ten trained seeds with replacement and recompute the PI-test RMSE of the resulting ensemble-mean prediction. The 95% seed-bootstrap intervals are $[0.2265, 0.2300]^\circ\text{C}$, $[0.2981, 0.3025]^\circ\text{C}$, and $[0.3345, 0.3445]^\circ\text{C}$ at 1-, 5-, and 10-year leads, respectively, around the reported ten-seed ensemble RMSEs of 0.2278, 0.3000, and 0.3394°C . Thus the PI ranking in Table 1 is insensitive to resampling the trained transition seeds.

The complete table that includes Table 1 is shown here as Table 2, including validation and test error on PI-control dataset, and generalization error in historical, early SSP and late SSP period.

Table 2: Prediction error across PI-control and forced-climate regimes. PI-control columns report validation and test error; forced columns report errors over the historical, early-warming (2015–2050), and late-warming (2070–2100) periods. Lower is better; the best model in each column is bold.

Model	Lead 1					Lead 5					Lead 10				
	PI-control		Forced			PI-control		Forced			PI-control		Forced		
	Val.	Test	Hist.	2015–50	2070–2100	Val.	Test	Hist.	2015–50	2070–2100	Val.	Test	Hist.	2015–50	2070–2100
L2	0.214	0.214	0.229	0.281	1.072	0.268	0.261	0.321	0.317	1.607	0.290	0.294	0.325	0.393	2.098
H1	0.213	0.214	0.229	0.287	1.103	0.269	0.261	0.320	0.320	1.704	0.290	0.294	0.325	0.399	2.132
POP	0.212	0.217	0.238	0.270	0.615	0.274	0.270	0.327	0.312	1.297	0.296	0.301	0.327	0.353	1.429
AI-Emulator	0.225	0.228	0.254	0.280	0.466	0.283	0.300	0.293	0.331	0.645	0.299	0.339	0.299	0.367	0.873

B Exact model specifications

All models receive the same five-channel PI-normalized state. They differ in what must be represented before the scalar SPG target is extracted.

B.1 Direct L^2 and H^1 predictors

For lead L , let X_L be the lead-specific design matrix normalized by per-feature mean and standard deviation, and let $Y_L \in \mathbb{R}^{n_L}$ denote the corresponding normalized scalar targets by the scalar mean and standard deviation.

$$X_{L,ti} = \frac{x_i(t) - \mu_i}{\sigma_{\text{PI},i}}, \quad Y_{L,t} = \frac{y(t) - \mu_y}{\sigma_{\text{PI},y}}$$

The direct predictors solve

$$\hat{m}_H = \arg \min_m \left\{ \frac{1}{n_L d} \|Y_L - X_L m\|_2^2 + m^\top H m \right\}, \quad (6)$$

The Euclidean predictor, i.e. L^2 model in the main text, uses $H = \alpha I$:

$$\hat{m}_{L^2} = \arg \min_m \left\{ \frac{1}{n_L d} \|Y_L - X_L m\|_2^2 + \alpha \|m\|_2^2 \right\}, \quad (7)$$

whereas the spatially regularized predictor, i.e. H^1 model uses $H = \alpha I + \gamma L_{\text{graph}}$:

$$\hat{m}_{H^1} = \arg \min_m \left\{ \frac{1}{n_L d} \|Y_L - X_L m\|_2^2 + \alpha \|m\|_2^2 + \gamma \|Dm\|_2^2 \right\}, \quad (8)$$

The predictor of the normalized variables $\tilde{x}$ and $\tilde{y}$ works as

$$\hat{y}_L = \hat{m}_M^\top \tilde{x}_L,$$

and the linear coefficient of predicting raw-physics target y from raw-physics input x with method M is

$$\hat{b}_{M,i} = \frac{\sigma_{\text{Pl},y}}{\sigma_{\text{Pl},i}} \hat{m}_{M,i}. \quad (9)$$

D is the discrete spatial-gradient operator and $L_{\text{graph}} = D^\top D$. The code does not explicitly construct the discrete gradient matrix D . Instead, it directly assembles the weighted graph Laplacian

$$L_{\text{graph}} = D^\top D$$

on the channel-specific ocean masks. The graph geometry follows the rotated grid described in Appendix A.3; in particular, all metric weights are computed using the rotated latitude rlat , rather than geographic latitude.

Let $a = (c, j, i)$ denote an active grid cell in channel c , rotated-latitude row j , and rotated-longitude column i . The graph connects only four-neighbor cells within the same channel:

$$(c, j, i) \leftrightarrow (c, j + 1, i)$$

in the rotated-latitude direction, and

$$(c, j, i) \leftrightarrow (c, j, i + 1)$$

in the rotated-longitude direction. No diagonal or cross-channel edges are included.

Conceptually, for an oriented edge $e = (a, b)$ with weight w_e , the corresponding row of D would be

$$D_{e,k} = \begin{cases} +\sqrt{w_e}, & k = a, \\ -\sqrt{w_e}, & k = b, \\ 0, & \text{otherwise.} \end{cases}$$

Thus

$$(Dm)_e = \sqrt{w_e}(m_a - m_b),$$

and the resulting quadratic penalty is

$$m^\top L_{\text{graph}} m = \|Dm\|_2^2 = \sum_{e \in E} w_e (m_a - m_b)^2.$$

The edge weights account for the rotated-grid metric. For an edge between adjacent rotated-latitude rows,

$$(c, j, i) \leftrightarrow (c, j + 1, i),$$

the code uses

$$w_{j+\frac{1}{2}}^{\text{rlat}} = \frac{\max(\cos \text{rlat}_j, 10^{-6}) + \max(\cos \text{rlat}_{j+1}, 10^{-6})}{2}.$$

For an edge along the rotated-longitude direction,

$$(c, j, i) \leftrightarrow (c, j, i + 1),$$

it uses

$$w_j^{\text{r lon}} = \frac{1}{\max(\cos \text{rlat}_j, 10^{-6})}.$$

An edge is included only when both neighboring cells are active in the mask for that channel. Hence, no edge crosses land, missing cells, the boundary of the north-of-40° analysis domain, or any other masked region.

Rather than constructing D and multiplying $D^\top D$, the implementation assembles L_{graph} edge by edge (we omit the subscript for simplicity). Each valid undirected edge $a \leftrightarrow b$ with weight w contributes

$$\begin{bmatrix} +w & -w \\ -w & +w \end{bmatrix}$$

to the corresponding a, b block:

$$L_{aa} += w, \quad L_{bb} += w, \quad L_{ab} -= w, \quad L_{ba} -= w.$$

Equivalently,

$$L_{ii} = \sum_{j: (i,j) \in E} w_{ij}, \quad L_{ij} = \begin{cases} -w_{ij}, & i \neq j \text{ and } i, j \text{ are connected neighbors,} \\ 0, & i \neq j \text{ otherwise.} \end{cases}$$

For the present dataset, the graph contains 16,678 undirected edges. The resulting matrix $L_{\text{graph}} \in \mathbb{R}^{9131 \times 9131}$ contains 42,487 stored nonzero entries and is symmetric positive semidefinite by construction.

B.2 POP/LIM low-rank state representation

POP first constructs its internal state coordinate: for raw-physics input column x ,

$$\tilde{x} = F(x - \mu_{\text{POP}}),$$

where μ_{POP} is the pixelwise mean in $\mathbb{R}^d$ and F is a diagonal matrix with diagonal element

$$f_i = \frac{\sqrt{\cos(\phi_i)}}{\sigma_{\text{PI}, c(i)}}.$$

Here ϕ_i is the geographic latitude and $\sigma_{\text{PI}, c(i)}$ is POP's pooled PI-training standard deviation for channel c . Notice that in POP we use per-channel standard deviation rather than per-feature as in L^2/H^1 . It then projects onto a rank-100 EOF basis: $z_t = U^\top \tilde{x}_t$, and fits the multivariate state evolution rather than the scalar SPG target directly. For a lead L , define

$$C(0) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} z_t z_t^\top, \quad C(L) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} z_{t+L} z_t^\top. \quad (10)$$

The lead-specific operator is estimated from these second moments, with ridge regularization in practice,

$$A_L \simeq C(L)C(0)^{-1}. \quad (11)$$

Table 3: Encoder architecture of the five-channel autoencoder.

Stage	Operation	Output per state
Input	normalized maps	$5 \times 99 \times 87$
enc1	ConvBlock, $5 \rightarrow 32$, stride 1	$32 \times 99 \times 87$
enc2	ConvBlock, $32 \rightarrow 64$, stride 2	$64 \times 50 \times 44$
enc3	ConvBlock, $64 \rightarrow 128$, stride 2	$128 \times 25 \times 22$
enc4	ConvBlock, $128 \rightarrow 256$, stride 2	$256 \times 13 \times 11$
enc5	ConvBlock, $256 \rightarrow 256$, stride 2	$256 \times 7 \times 6$
Bottleneck	flatten, linear $10752 \rightarrow 64$	64

Table 4: Decoder architecture. Resolution is restored by ordinary stride-one convolutional blocks followed by bilinear interpolation, not by forward transposed convolutions.

Stage	Operation	Output per state
Latent	input	64
Projection	linear $64 \rightarrow 10752$, reshape	$256 \times 7 \times 6$
dec4	ConvBlock $256 \rightarrow 256$; interpolate	$256 \times 13 \times 11$
dec3	ConvBlock $256 \rightarrow 128$; interpolate	$128 \times 25 \times 22$
dec2	ConvBlock $128 \rightarrow 64$; interpolate	$64 \times 50 \times 44$
dec1	ConvBlock $64 \rightarrow 32$; interpolate	$32 \times 99 \times 87$
Output	1×1 convolution $32 \rightarrow 5$	$5 \times 99 \times 87$

A separately fitted lead-specific readout c_L with linear regression gives the prediction

$$\frac{\hat{y}_{t+L} - \mu_{y,L}}{\sigma_{y,L}} = c_L^\top A_L z_t, \quad (12)$$

This factorization imposes two forms of structure absent from direct scalar regression: a hard low-rank state subspace and a linear state-evolution constraint within that subspace. The linear coefficient of predicting raw-physics target y from raw-physics input x is (written as column vector)

$$\hat{b}_{\text{POPL}} = \sigma_{y,L} F U \hat{A}_L^\top \hat{c}_L \quad (13)$$

B.3 Convolutional autoencoder

The learned representation uses a frozen convolutional autoencoder with latent dimension 64. Every convolutional block has the form

$$\text{Conv}_{3 \times 3} \rightarrow \text{BatchNorm} \rightarrow \text{SiLU} \rightarrow \text{Conv}_{3 \times 3} \rightarrow \text{BatchNorm} \rightarrow \text{SiLU}.$$

The first convolution in each block uses the listed stride; the second uses stride one, with padding one throughout. The encoder contains 5,506,309 trainable parameters in the implementation.

With encoder E_ϕ and decoder D_ϕ , we train it by reconstructing the per-channel normalized state:

$$a_i(t) = \frac{x_i(t) - \mu_i}{\sigma_{PI,c(i)}}. \quad (14)$$

$$z_t = E_\phi(a_t) \in \mathbb{R}^{64}, \quad \hat{a}_t = D_\phi(z_t).$$

The autoencoder is trained with masked, rotated-latitude-weighted reconstruction MSE,

$$\mathcal{L}_{\text{AE}}(\phi) = \frac{\sum_{t \in \mathcal{B}} \sum_c \sum_{ij} M_{cij} g_{ij} [D_\phi(E_\phi(a_t))_{cij} - a_{tcij}]^2}{\sum_{t \in \mathcal{B}} \sum_c \sum_{ij} M_{cij} g_{ij}}, \quad (15)$$

where

$$g_{ij} = \frac{\max(w_{ij}, 0)}{\text{mean}_{kl}[\max(w_{kl}, 0)]}.$$

Training uses AdamW with learning rate 10^{-3} and weight decay 10^{-6} . The trained encoder and decoder are frozen before transition training.

B.4 Latent transition and recursive rollout

The raw latent archive is standardized dimension-by-dimension using PI training years 1–1400,

$$q_{t,j} = \frac{z_{t,j} - \mu_j^{\text{train}}}{\sigma_j^{\text{train}}}, \quad j = 1, \dots, 64. \quad (16)$$

One time-homogeneous residual transition is shared across all annual rollout steps,

$$\hat{q}_{k+1} = F_\theta(\hat{q}_k) = \hat{q}_k + f_\theta(\hat{q}_k), \quad (17)$$

where f_θ has widths $64 \rightarrow 256 \rightarrow 256 \rightarrow 256 \rightarrow 64$, SiLU after each hidden linear layer, zero dropout, and a zero-initialized final linear layer. The transition contains 164,672 trainable parameters in the implementation.

Starting from the encoded truth q_t , the transition is trained recursively over a horizon n with

$$\mathcal{L}_{\text{lat}}(\theta) = \frac{1}{n} \sum_{j=1}^n c_j \left[\frac{1}{B d_z} \sum_{b=1}^B \|\hat{q}_{b,t+j} - q_{b,t+j}\|_2^2 \right], \quad c_j = \frac{j^{-1/2}}{n^{-1} \sum_{r=1}^n r^{-1/2}}, \quad (18)$$

with $d_z = 64$. AdamW learning rate 3×10^{-4} , weight decay 10^{-5} , batch size 128, gradient-norm clipping at one, at most 300 epochs, validation every five epochs, and patience 20.

The inference chain is

$$\begin{aligned} x_t &\xrightarrow{\text{normalization}} a_t \xrightarrow[\text{frozen}]{E_\phi} z_t \xrightarrow[\text{PI statistics}]{\text{standardize}} q_t \xrightarrow[\text{recursive}]{F_\theta} \hat{q}_{t+1:t+10} \\ &\xrightarrow[\text{frozen}]{\text{unscale}+D_\phi} \hat{a}_{t+1:t+10} \rightarrow \hat{x}_{t+1:t+10} \xrightarrow{\text{SPG mean}} \hat{y}_{t+1:t+10}. \end{aligned}$$

In this study, we use the averages of the predictions of ten independently trained transitions with different random seeds.

C Model selection and complete prediction results

C.1 Validation-only selection

For L^2 model in this study $\alpha = 1e - 3$, and for H^1 model $\alpha = 1e - 3$ and $\gamma = 1e - 3$. These primary hyperparameters are selected using the equal-weight mean of the complete PI validation RMSEs at leads 1, 5, and 10.

C.2 POP rank selection and ridge sweep

In POP implementation, we set $\text{rank}r = 100$. The rank choice is made to be able to almost fully reproduce the global-warming state (projecting on first 50 EOF causes a cold bias, not shown). Eq. (11) and Eq. (12) show the unregularized POP form. In implementation, we stabilize A_L with a small ridge $\rho = 1e - 2$ to improve the conditioning of the inverse, as in Eq. (19):

$$A_L(\rho)^\top = (Z_-^\top Z_- + \alpha_\rho I)^{-1} Z_-^\top Z_{+L}, \quad \alpha_\rho = \rho \frac{\text{tr}(Z_-^\top Z_-)}{r}. \quad (19)$$

The difference between exact POP when $\rho = 0$ and when $\rho = 1e - 2$ is less than 0.5% on validation set. However, we find if we continue increasing ρ , the prediction in the forcing climate will further approach to L^2 (see Fig. 4). This behavior is understandable by viewing POP as an increasingly strongly ridge-shrunk latent predictor as ρ increases. When $\rho \rightarrow \infty$, $A_L \rightarrow 0$ and the prediction converges to climatology.

C.3 Latent-dynamics roll-out

We selected a two-step rollout loss $n = 2$ in Eq. (18) as a short-horizon compromise that reduced recursive cold drift relative to one-step training while limiting the variance damping associated with longer rollout objectives.

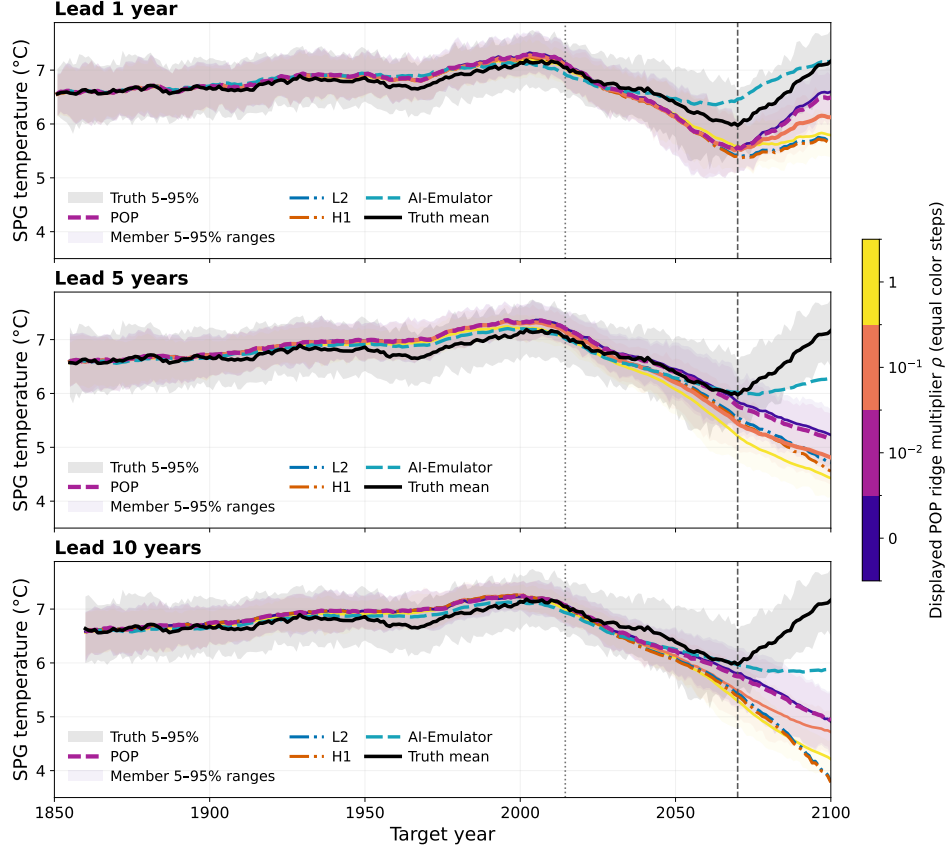


Figure 4: POP prediction across the archived latent-operator ridge sweep in historical and SSP 3-7.0 climate. L^2 , H^1 , and the nonlinear emulator are shown as references.

C.4 Forced-trained H^1 diagnostic reference

To aid physical interpretation of the late-climate attribution, we fit an additional H^1 regression using forced-climate training data through 1850-2090. This model is not included in the transfer comparison of Table 1: unlike the four primary predictors, it has access to the forced regime and therefore serves only as a diagnostic reference for the response patterns available when forced-climate information is observed. The prediction result is shown in Fig. 5

The model uses the same state construction and H^1 objective as the PI-trained predictor, with $\alpha = 1e-3$ and $\gamma = 1e-2$. $\gamma = 1e-3$ produces almost identical prediction curves and marginally better RMSE (<1% on average of 1-, 5- and 10- leads), but noticeably more spatially irregular sensitivity fields. The broad sensitivity structure remains, while local grid-scale signs and amplitudes become less stable. $\gamma = 1e-2$ is therefore preferable for smooth regional interpretation, but not because it predicts better.

We apply the same finite-change attribution defined in Appendix D and show the resulting channel and regional contributions alongside the PI-trained models in Fig. 3 and Appendix E.

D Finite late-climate contribution attribution

The main-text heatmap uses a finite-change attribution in late-warming period. The estimand asks which input channels and regions contribute to each model's predicted change across the late-warming period.

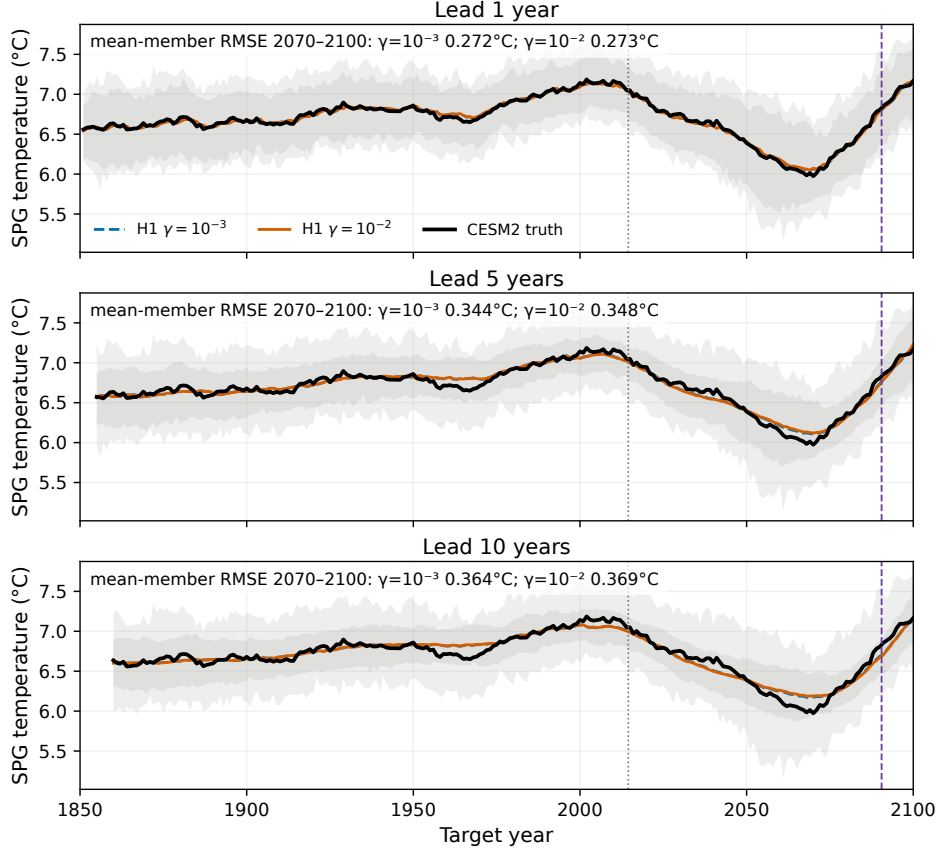


Figure 5: Effect of the H^1 spatial-smoothing strength on forced-climate prediction. Ensemble-mean SPG temperature predictions from the forced-only H^1 models with $\gamma = 10^{-3}$ (blue dashed) and $\gamma = 10^{-2}$ (orange) are compared with the CESM2 ensemble mean (black) at 1-, 5-, and 10-year lead times. Shading denotes the corresponding inter-member spread. Both models use the same 70 training members and data through 2090, with $\alpha = 10^{-3}$, and no explicit CO_2 predictor; vertical lines mark 2015 and 2090. The reported values are mean memberwise RMSEs over 2070–2100. Reducing γ produces nearly unchanged ensemble-mean trajectories and a small RMSE improvement, while permitting substantially less spatially smooth coefficient fields.

D.1 Endpoint-window contrast

Let p index one of the 100 forced members, i an input feature, and $L \in \{1, 5, 10\}$ the forecast lead. Target years label all windows, so the source state for target year t is year $t - L$. Define the endpoint states of raw physical input quantities $x_{p,t,i}$

$$\bar{x}_{piL}^{\text{first}} = \frac{1}{5} \sum_{t=2070}^{2074} x_{p,t-L,i}, \quad (20)$$

$$\bar{x}_{piL}^{\text{last}} = \frac{1}{5} \sum_{t=2096}^{2100} x_{p,t-L,i}. \quad (21)$$

The window midpoints, 2072 and 2098, are separated by 26 yr, or $\Delta\tau = 2.6$ decades. The aligned late-period finite-difference response rate is per-decade

$$\dot{x}_{piL} = \frac{\bar{x}_{piL}^{\text{last}} - \bar{x}_{piL}^{\text{first}}}{2.6}. \quad (22)$$

D.2 L^2 and H^1 contributions

Let $b_i^{M,L}$ be the effective coefficient of direct model $M \in \{L^2, H^1\}$ in physical input and target units in Eq. (9). The pointwise late-response contribution for member p is

$$\chi_{pi}^{M,L} = b_i^{M,L} \dot{x}_{piL}. \quad (23)$$

Centering and intercept terms cancel in the endpoint difference, so

$$\sum_i \chi_{pi}^{M,L}$$

is exactly the model’s predicted endpoint-window response rate for that member.

D.3 POP contributions

Following Eq. 23, the pointwise contribution for member p is

$$\chi_{pi}^{\text{POP},L} = b_{\text{POP},L,i} \dot{x}_{piL}, \quad (24)$$

with $b_{\text{POP},L,i}$ in Eq. 13 and therefore has the same exact additivity as the direct linear models.

D.4 Nonlinear integrated-gradient contribution

For the nonlinear model, attribution is evaluated in the PI-standardized physical-state coordinate a in Eq. 14. For each lead, a_L^0 and a_L^1 are the ensemble-mean source states corresponding to target windows 2070–2074 and 2096–2100, respectively. Let $F_{s,L}(a)$ be the scalar SPG prediction from seed s . The contrastive integrated-gradient contribution of feature i is

$$\text{IG}_i^{s,L} = (a_{i,L}^1 - a_{i,L}^0) \int_0^1 \frac{\partial F_{s,L}}{\partial a_i} [a_L^0 + \alpha(a_L^1 - a_L^0)] d\alpha. \quad (25)$$

The implementation uses a 32-step midpoint rule, $\alpha_k = (k - \frac{1}{2})/32$. The neural integrated-gradient was computed by PyTorch backpropagating through the complete computational graph and reports

$$\chi_i^{\text{AI},s,L} = \frac{\text{IG}_i^{s,L}}{2.6}. \quad (26)$$

The Jacobians along the linear interpolation path are generally close to each other; the cancellation mainly happens between ten independently random seeds, especially lead 10 (not shown).

Integrated-gradient completeness gives

$$\sum_i \chi_i^{\text{AI},s,L} \simeq \frac{F_{s,L}(a_L^1) - F_{s,L}(a_L^0)}{2.6}, \quad (27)$$

up to quadrature error.

D.5 Channel totals and normalization used in the main heatmap

For channel c ,

$$C_c^{M,L} = \sum_{i:K(i)=c} \chi_i^{M,L}, \quad (28)$$

where $K(i)$ maps compact feature i to one of the five input channels. The physical quantity $C_c^{M,L}$ has units of $^\circ\text{C decade}^{-1}$ and retains the cancellations between positive and negative feature contributions.

The main-text Figure 3 displays a dimensionless normalized signed contribution,

$$\tilde{C}_c^{M,L} = \frac{C_c^{M,L}}{\sum_{c'} |C_{c'}^{M,L}|}, \quad (29)$$

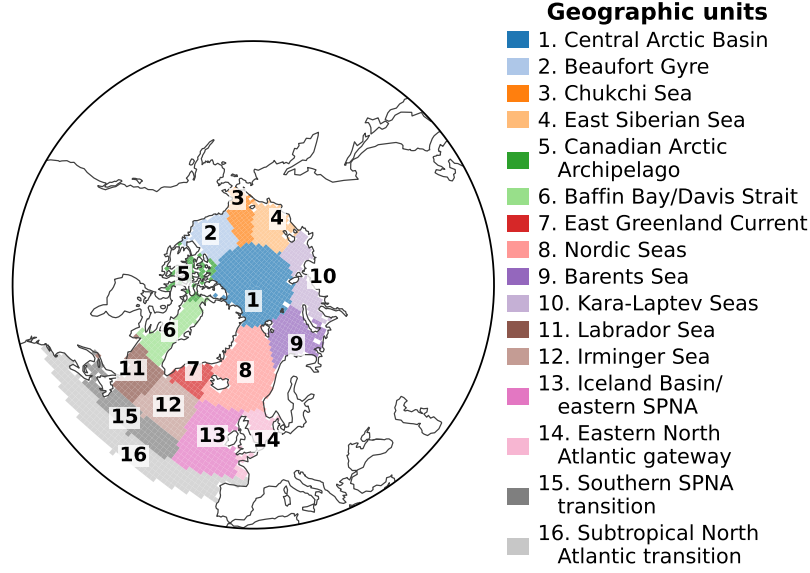


Figure 6: Non-overlapping partition in Tab. 5

E Geographic aggregation and complete attribution atlas

E.1 Non-overlapping North40 partition

Every active North40 surface-ocean cell is assigned to exactly one of 16 geographic units by applying the candidate rules in Table 5 in precedence order (shown in Figure 6). Once an earlier rule claims a cell, that cell is excluded from all later regions. This gives a mutually exclusive and exhaustive partition over the 2,095 active surface-ocean cells. Subsurface channels contain fewer valid compact features because their masks differ.

Table 5: Precedence-resolved geographic units used for regional attribution. Longitude is represented on $[-180^\circ, 180^\circ)$; lower bounds are inclusive and upper bounds exclusive unless stated otherwise.

No.	Region	Candidate rule before precedence
1	Central Arctic Basin	$\text{latitude} \geq 80^\circ$
2	Beaufort Gyre	$66.5^\circ \leq \text{lat} < 80^\circ, -160^\circ \leq \text{lon} < -120^\circ$
3	Chukchi Sea	$65^\circ \leq \text{lat} < 80^\circ, -180^\circ \leq \text{lon} < -160^\circ$
4	East Siberian Sea	$68^\circ \leq \text{lat} < 80^\circ, 140^\circ \leq \text{lon} < 180^\circ$
5	Canadian Arctic Archipelago	$68^\circ \leq \text{lat} < 80^\circ, -120^\circ \leq \text{lon} < -75^\circ$
6	Baffin Bay/Davis Strait	$60^\circ \leq \text{lat} < 80^\circ, -75^\circ \leq \text{lon} < -50^\circ$
7	East Greenland Current	$60^\circ \leq \text{lat} < 80^\circ, -50^\circ \leq \text{lon} < -25^\circ$
8	Nordic Seas	$60^\circ \leq \text{lat} < 80^\circ, -25^\circ \leq \text{lon} < 15^\circ$
9	Barents Sea	$65^\circ \leq \text{lat} < 80^\circ, 15^\circ \leq \text{lon} < 60^\circ$
10	Kara-Laptev Seas	$68^\circ \leq \text{lat} < 80^\circ, 60^\circ \leq \text{lon} < 140^\circ$
11	Labrador Sea	$50^\circ \leq \text{lat} < 68^\circ, -70^\circ \leq \text{lon} < -45^\circ$
12	Irminger Sea	$50^\circ \leq \text{lat} < 68^\circ, -45^\circ \leq \text{lon} < -25^\circ$
13	Iceland Basin/eastern SPNA	$45^\circ \leq \text{lat} < 65^\circ, -25^\circ \leq \text{lon} < -5^\circ$
14	Eastern North Atlantic gateway	$45^\circ \leq \text{lat} < 65^\circ, -5^\circ \leq \text{lon} < 20^\circ$
15	Southern SPNA transition	$45^\circ \leq \text{lat} < 50^\circ, -70^\circ \leq \text{lon} < -25^\circ$

No.	Region	Candidate rule before precedence
16	Subtropical North Atlantic transition	$40^\circ \leq \text{lat} < 45^\circ, -80^\circ \leq \text{lon} < 20^\circ$

E.2 Region–channel contributions

Let $R(i)$ and $K(i)$ be the unique region and channel of feature i . For a classical forced member or neural seed, generically indexed by u ,

$$C_{u,r,c}^{M,L} = \sum_i \mathbf{1}\{R(i) = r\} \mathbf{1}\{K(i) = c\} \chi_{ui}^{M,L}. \quad (30)$$

The displayed quantity is therefore the *net contribution* of that region–channel block to the scalar SPG prediction.

For L^2 , H^1 , and POP, Eq. (30) is evaluated separately for all 100 forced members and then averaged. For the nonlinear model it is evaluated for each of the ten trained transitions along the common ensemble-mean contrast and then averaged over seeds.

E.3 Dominant lead-10 region–channel blocks

The current Fig. 3 displays the following eight lead-10 blocks: subtropical $T_{200-1000}$, southern-SPNA SSH, subtropical $S_{200-1000}$, Labrador SSH, southern-SPNA $S_{200-1000}$, Barents $S_{200-1000}$, Nordic-Seas $S_{200-1000}$, and Kara–Laptev $S_{200-1000}$. The qualitative distinction is that POP retains a large positive subtropical/southern subsurface-temperature pathway, whereas the nonlinear emulator places much more of its positive response in salinity, including a positive Nordic-Seas subsurface-salinity contribution.

The eight blocks are selected before normalization from all $16 \times 5 = 80$ possible region–channel pairs at lead 10. For each pair (r, c) , we calculate

$$\text{score}_{r,c} = \max_{m \in \{L2, H1, POP, AI\}} |C_{m,r,c}|.$$

and then select the eight blocks with the highest score. Here $C_{m,r,c}$ is the raw contribution for model m , region r , and channel c , in $^\circ\text{C decade}^{-1}$. The normalization in Fig. 3 happens only after these eight blocks have been selected:

$$\tilde{C}_{m,r,c} = \frac{C_{m,r,c}}{\sum_{(r',c') \in \mathcal{P}} |C_{m,r',c'}|},$$

where $\mathcal{P}$ is this fixed set of eight pairs. Thus, normalization changes the colors but not which blocks are included or their order. H1 (forced) is excluded from the selection score, and is displayed afterward for the same eight blocks as a diagnostic reference.

E.4 Complete polar atlas

The complete atlas uses 25 panels per lead: five model rows and five channel columns. Each regional scalar is painted uniformly over the region’s common surface footprint for visualization.

F Expanded forced-climate training support with CO₂ encoding

We perform auxiliary experiments in which forced-climate samples are added to the PI training set. Unlike the primary zero-adaptation experiment, these predictors receive both the ocean-state map and an encoding of the prescribed atmospheric CO₂ trajectory. We use a fixed 70/15/15 forced-member split: 70 members for fitting, 15 for validation where applicable, and 15 for final testing.

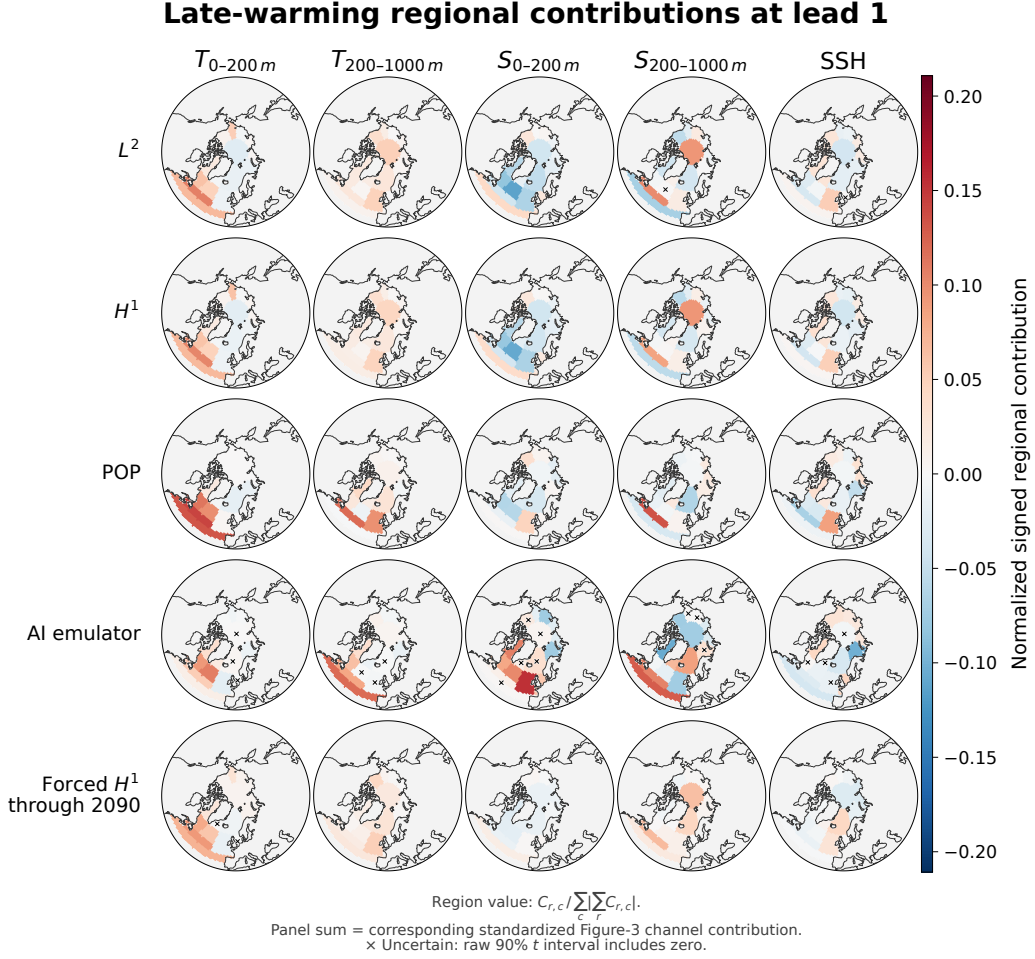


Figure 7: Regional and channel decomposition of the late-warming prediction. Rows show L^2 , H^1 , POP, the AI emulator, and the forced H^1 model trained through 2090; columns show the five ocean-state channels. For each predictor and lead, the raw regional contribution $\bar{C}_{r,c}$ is expressed in $^{\circ}\text{C decade}^{-1}$ and normalized as $\tilde{C}_{r,c} = \frac{\bar{C}_{r,c}}{\sum_c |\sum_r \bar{C}_{r,c}|}$. Colors therefore show the signed geographical distribution relative to the total absolute contribution of the five channels, rather than absolute prediction magnitude. Raw contributions are calculated from coefficient–input–trend products for the linear predictors and path-integrated Jacobian products for the AI emulator. A black cross indicates that the raw two-sided 90% t -interval contains zero. The corresponding signed mean remains displayed and is retained in all channel and regional sums. No spatial smoothing or interpolation is applied.

Table 6: Training-only CO₂ coordinate constants for the PI+historical fit.

Lead L (yr)	κ_L	$s_{f,L}$ (ppm)	$s_{r,L}$ (ppm)
1	0.021499	39.874873	0.333405
5	0.112717	37.003744	1.129663
10	0.239192	33.676799	2.066245

F.1 CO₂ encoding

When forced-climate samples are included in training, we augment the ocean state with a compact encoding of the prescribed CO₂ trajectory, allowing the predictor to distinguish climate states associated with different forcing levels. Let C_t be the annual atmospheric CO₂ concentration in ppm, formed by concatenating historical forcing through 2014 with SSP3–7.0 thereafter. The PI-control CO₂ concentration is $C_{PI} = 284.7$ ppm. For a source year t and lead L , we encode both the relative amplitude of CO₂ concentration and the increment of CO₂ in the input,

$$f_t = C_t - C_{PI}, \quad h_{t,L} = C_{t+L} - C_t. \quad (31)$$

For normalization, the two columns are orthogonalized and scaled using training rows only (Gram-Schmidt):

$$\kappa_L = \frac{f^\top h_L}{f^\top f}, \quad r_L = h_L - \kappa_L f, \quad (32)$$

$$s_{f,L} = \left(\frac{1}{n_L} f^\top f \right)^{1/2}, \quad s_{r,L} = \left(\frac{1}{n_L} r_L^\top r_L \right)^{1/2}, \quad (33)$$

and the forcing coordinate is

$$u_{t,L} = \begin{bmatrix} q_{f,t,L} \\ q_{r,t,L} \end{bmatrix} = \begin{bmatrix} f_t/s_{f,L} \\ (h_{t,L} - \kappa_L f_t)/s_{r,L} \end{bmatrix}. \quad (34)$$

There is no centering of the forcing columns, so $u_{t,L} = 0$ for every PI row. Orthogonalization is only a numerical reparameterization of the span of the raw forcing columns.

F.2 State-plus-CO₂ direct regression

With X_L and Y_L defined exactly as in B.1 and with joint objective similar to Eq 1 and Eq. 6, the loss function including $u_{t,L}$ as additional variable is

$$\min_{m_L, \gamma_L} \frac{1}{n_L d} \|Y_L - X_L m_L - Q_L \beta_L\|_2^2 + m_L^\top H m_L, \quad (35)$$

where $H = \alpha I$ for L^2 and $H = \alpha I + \gamma D^\top D$ for H^1 , with same hyperparameter as L^2 and H^1 in implementation. Q_L = matrix whose rows are $u_{t,L}^\top$, $\beta_L \in \mathbb{R}^2$. The two forcing coefficients are unpenalized. Thus CO₂ is not added after fitting a frozen state-only predictor; the state and forcing coefficients are estimated jointly. The predictor is

$$\hat{y}_{t+L} = \tilde{x}_t^\top m_L + u_{t,L}^\top \beta_L.$$

F.3 Conditioned direct POP

In the CO₂ forced POP, CO₂ is not injected into the latent dynamics. We still took the same hyperparameter as unforced POP. After fitting A_L , the latent readout and forcing coefficients are fitted jointly:

$$\min_{c_L, \tilde{c}_L} \|Y_L - (Z_L A_L^\top) c_L - Q_L \beta_L\|_2^2, \quad (36)$$

so the physical forecast is

$$\hat{y}_{t+L} = \mu_{y,L} + \sigma_{y,L} (c_L^\top A_L z_t + \beta_L^\top u_{t,L}). \quad (37)$$

The explicit forcing is therefore an output-level scenario correction, while historical climate information may also enter implicitly through the state basis and fitted operator.

F.4 Expanded support and the incremental effect of CO₂ encoding

Fig. 10 compares predictions as forced-climate training support is expanded beyond PI control with CO₂ encoded as well. Expanded training support and CO₂ encoding improve some forecasts, particularly before the late-century reorganization, but the benefit is not universal and the models still fail to reproduce the late-warming response. This experiment is intended as an exploratory control rather than an optimized treatment of external forcing. Its purpose is to distinguish lack of scenario information from failure to transfer across a changed dynamical regime. Scalar CO₂ encoding with expanding historical data support can compensate for missing scenario information in some support–lead combinations, most clearly for the longer-lead direct regressions trained only through the historical period, but it does not resolve the post-reorganization extrapolation error or remove the dependence of transfer skill on model representation.

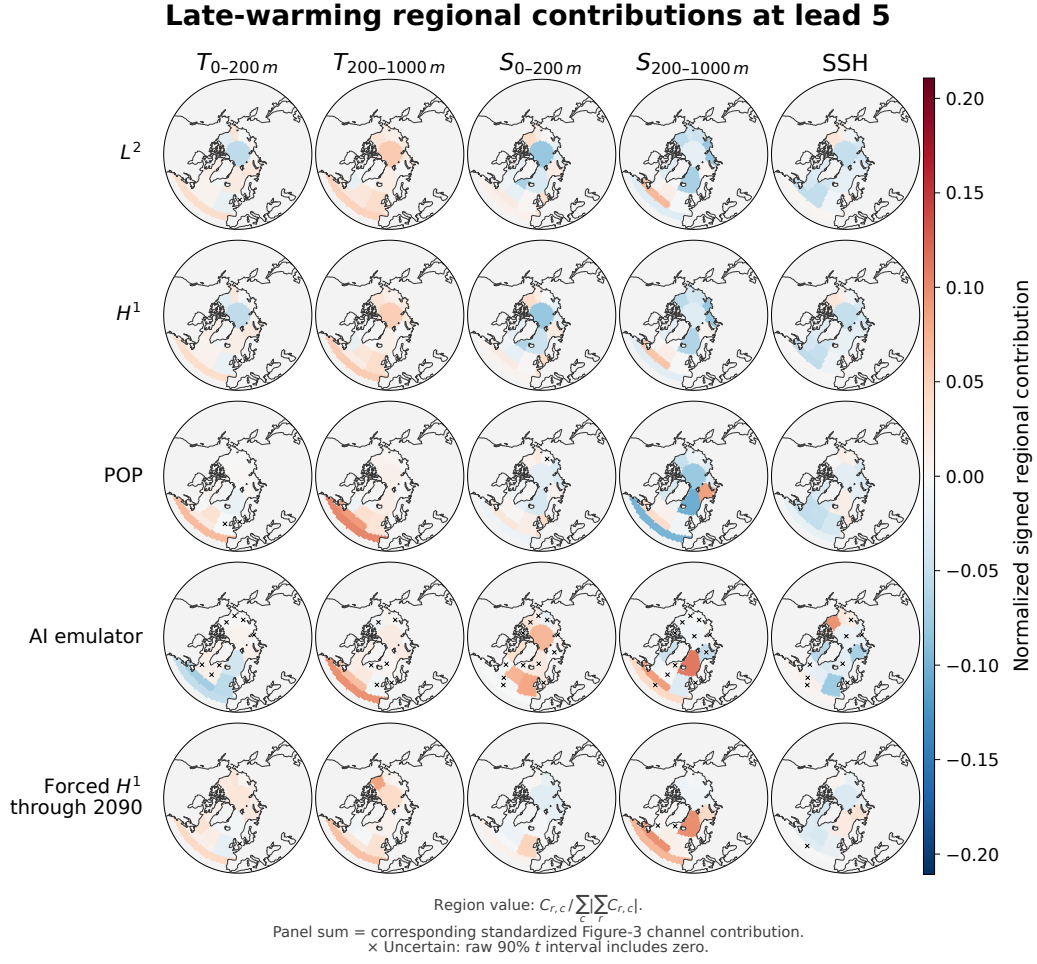


Figure 8: Same as Figure 7, but for lead = 5

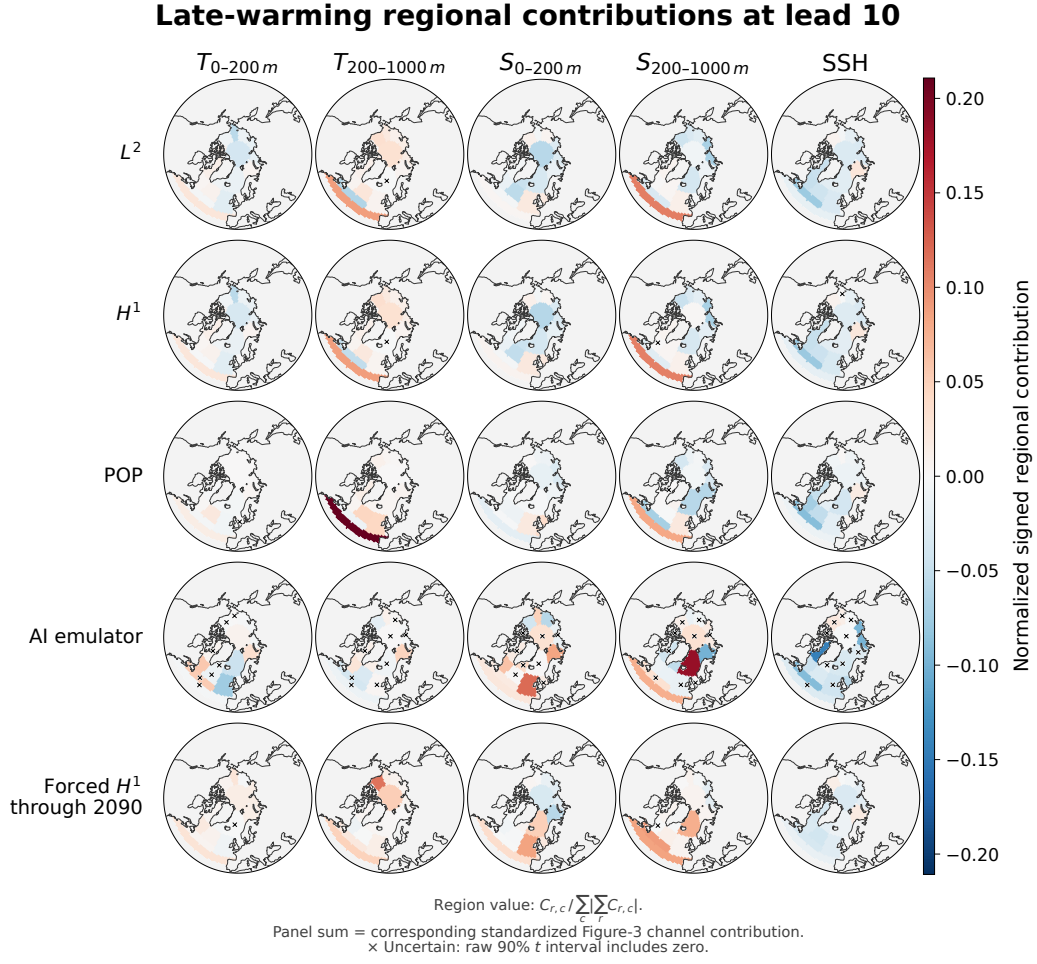


Figure 9: Same as Figure 7, but for lead = 1

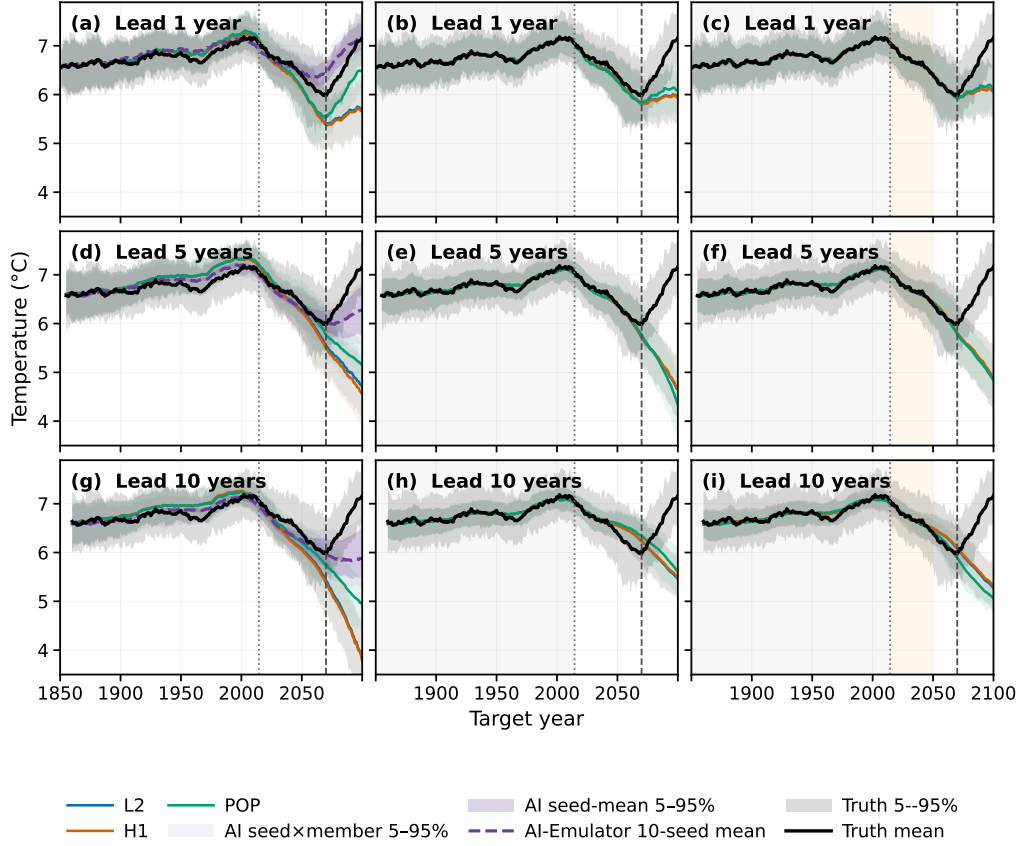


Figure 10: Historical and SSP3-7.0 ensemble predictions under progressively expanded training support. Rows show forecast leads of 1, 5, and 10 years. Columns correspond to models trained on PI control only (left), PI control plus historical simulations through 2014 (center), and PI control plus historical and early-SSP simulations through 2049 (right). Gray and pale-amber backgrounds mark the historical and early-SSP periods included in training, respectively. Colored curves show the 100-member ensemble-mean predictions from L2, H1, and POP, with bands denoting their 5th–95th percentile ranges. In the left column, the dashed purple curve shows the 10-seed AI-emulator mean; the purple bands represent pooled seed–member and seed-mean uncertainty. Black curves and gray bands show the simulated ensemble mean and 5th–95th percentile range. The dotted vertical line marks the historical–SSP transition, and the dashed line marks the late-century analysis boundary at 2070.