Architectural Sensitivity of AI Weather Prediction Models to 3D Structural and Seasonal Climate Forcing

Mozhgan Amiramjadi[1], Christopher Roth[1], Peer Nowack[1,2]

mozhgan.amjadi@kit.edu

[1]Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

[2]Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Germany

Objective & Experimental Design

Core Hypothesis:

State-of-the-art AI weather models exhibit “climatological memory”, causing systematic drift toward their training distribution when initialized under Out-of-Distribution (OOD) forcing.

Key Diagnostic Question:

How does architecture, specifically the transition from purely data-driven ML models to hybrid Physics-ML models, influence the preservation of physical consistency–such as the vertical structure of key atmospheric variables–under these shifting climate states?

Data:

- Historical Baseline (ERA5, 1955)
- Modern Era/Control (ERA5, 2023)
- Future Climate Scenario (nextGEMS, 2049)

The Models:

- Deterministic: NeuralGCM (Hybrid), GraphCast (GNN), AIFS (Transformer).
- Probabilistic: GenCast (Conditional Diffusion).

	GraphCast	GenCast	AIFS	NeuralGCM
Training period	1979–2017 (ERA5)	1979–2018 (ERA5)	1979–2020 (ERA5 + Op. Analysis)	1979–2019 (ERA5)
Number of pressure levels (hPa range)	13 (1000–50)	13 (1000–50)	13 (1000–50)	37 (1000–1)
Horizontal resolution	1°	1°	0.25° (N320)	0.7° / 1.4° / 2.8°

Table 1: Model Configurations. Summary of training periods, horizontal resolutions, and vertical levels for the four architectures. The list reflects our specific configuration (not the full capability of each model).

Resolution Sensitivity in Hybrid Atmospheric Architectures

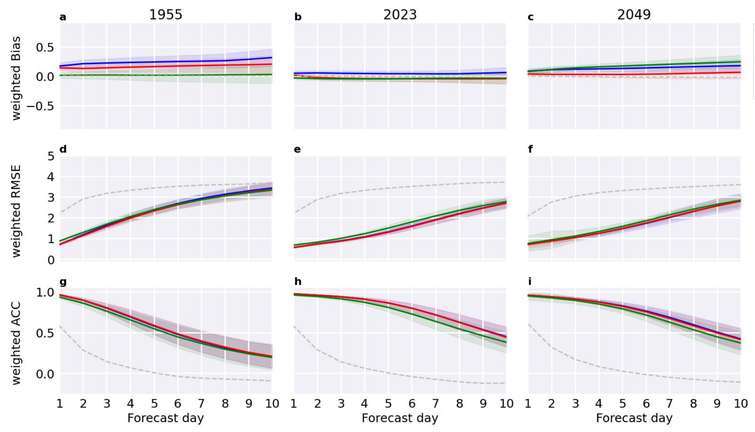


Figure 1: Global forecast evaluation for temperature (K) at 1000 hPa using NeuralGCM. Panels show weighted bias (top), RMSE (middle), and ACC (bottom) for the 1955, 2023, and 2049 climate states.

Global Skill Robustness:
Over 1–10 day leads, global predictive skill (ACC/RMSE) remains remarkably stable across different resolutions.

The Drift “Sweet Spot”:
The 1.4° configuration demonstrated the highest structural stability, effectively minimizing mean systematic drift compared to coarser or finer setups.

Robustness to Climate Shift:
All three resolution scales maintain skill significantly above persistence, demonstrating that horizontal scaling alone does not drastically alter the hybrid model's fundamental response to OOD climate states at the surface.

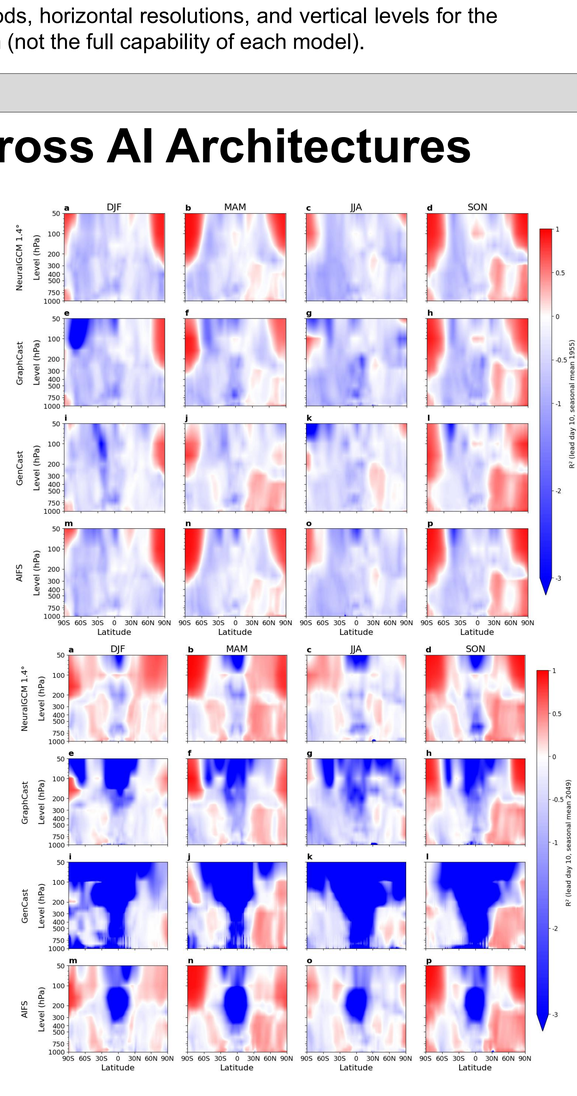
Structural Robustness Across AI Architectures

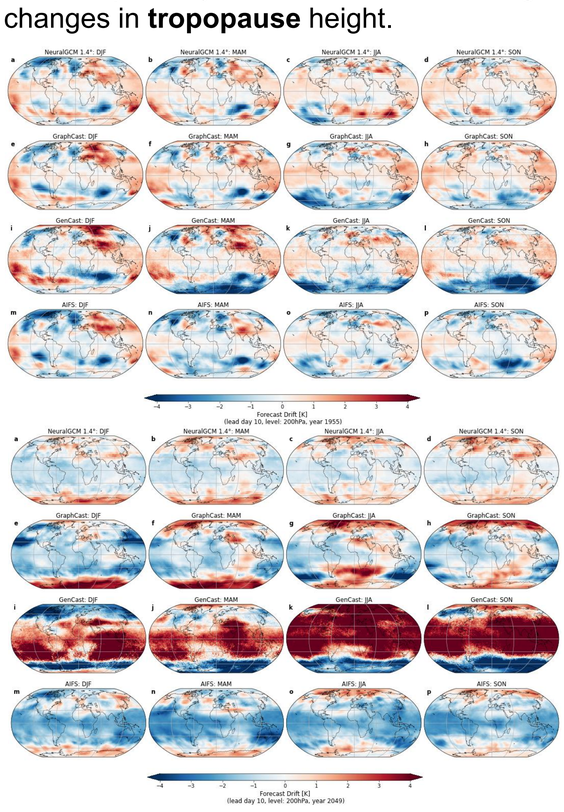
Vertical Structural Stability (R^2)

Regime-Dependent Skill:
There are seasonally robust patterns of high R^2, e.g. 60-90N SON, linked to high intrinsic predictability. However, we find strong reductions in predictive skill for several high latitude and altitude regions, suggesting that AI models (architecture-dependent) could struggle with non-linear dynamic shifts (e.g., polar vortex) when pushed outside their training climatology.

The Stratospheric “Cliff”:
Predictive power (R^2) effectively vanishes above 200 hPa in the OOD climates. This “blue-out” is most severe in 13-level models (**GraphCast, GenCast, AIFS**).

Vertical Resolution Advantage:
NeuralGCM’s 37-level architecture better maintains skill across the atmospheric column, possibly linked to better resolving changes in **tropopause** height.



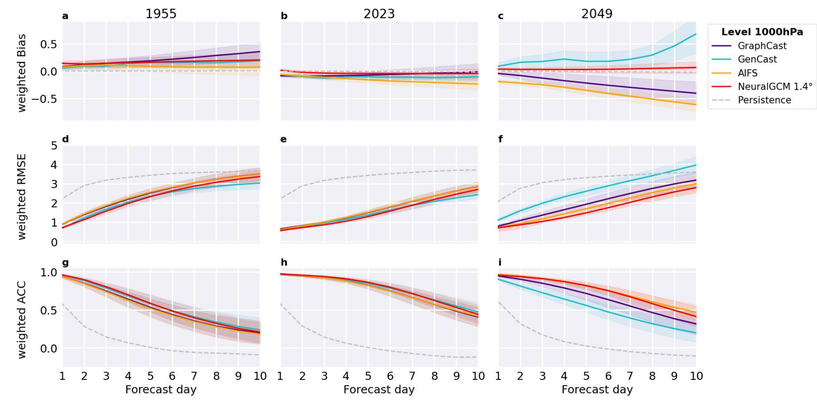


Climatological Drift & Dynamical Memory

Historical “Warm-Drift” (1955):
Models forced with cooler 1955 states exhibit a systematic warm bias, attempting to “pull” the atmosphere back toward their training climatology.

Future “Climatological Drag” (2049):
In the 2049 scenario, large-scale drifts align with the expected climate-shift signature; models struggle to maintain the “new” warm state, drifting back toward the cooler training distribution.

Hybrid Anchoring:
NeuralGCM shows the least drift, as its differentiable dynamical core acts as a physical constraint.



Surface-Level Skill & Bias Stability

Bias Stability:
The models exhibit a stable bias profile in the 2023 “Control” era. However, in the 2049 “Future” state, we see a significant divergence. AIFS (orange) and GraphCast (indigo) represent a negative bias trend and GenCast (cyan) shows a sharp positive bias over the 10-day lead time, while NeuralGCM (red) remain more centered.

RMSE Growth:
RMSE trajectories are consistent across 1955, 2023, and 2049 for all models. This suggests that while mean states shift, the rate of error growth at the surface is primarily governed by chaotic atmospheric limits rather than climatological forcing.

ACC Persistence:
All models maintain high Anomaly Correlation Coefficients (>0.6) through Day 5-6 across all eras. NeuralGCM and AIFS show slightly higher skill retention at Day 10 compared to GraphCast, particularly in the 2049 scenario.

Conclusion:
Physical Adaptation vs. Statistical Compensation

- ✓ **The Error-Optimal Trap:** Models optimized to minimize error on historical data (e.g., GraphCast) often drift toward training period for the global error, but regional biases can diverge (surface and higher altitudes).
- ✓ **Dynamical Core & Forcing:** Unlike purely data-driven models, NeuralGCM uses a differentiable dynamical core to solve large-scale equations. Combined with prescribed forcings, this allows the model to physically adjust to OOD states rather than “correcting” them back to the training climatological mean.
- ✓ **Vertical Anchor:** Resolving the Cold-Point Tropopause is a direct result of NeuralGCM’s 37-level discretization and its physics-based solver, preventing the vertical skill collapse (blue-out) seen in 13-level architectures.
- ✓ **Regime-Dependent Reliability:** AI skill remains high in stable regimes but collapses during non-linear transitions (e.g., SSW). A hybrid approach is helpful to prevent catastrophic climatological drift in a warming world.