LOW-LOSS TEMPORAL CONNECTIVITY IN WEATHER FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI weather models are usually trained for one fixed forecast step: a 6-hour model outputs only 6-hour forecasts. To obtain a forecast at an in-between horizon such as 4 hours, one currently has to retrain a new model, chain several short-step models, or interpolate their outputs on the grid. Each option is costly, blurs fine spatial detail, or both. This is a real obstacle for downstream systems (ocean, sea-ice, hydrology and air-quality models, and sub-6-hour data assimilation) that need forcing at cadences unrelated to 6 hours. We show empirically that intermediate horizons can instead be recovered directly in a model's weights. Starting from a public 6-hour global model checkpoint (Aurora), we fine-tune 3-hour and 1-hour versions with an identical recipe. Comparing internal representations with Centered Kernel Alignment (CKA), the 3-hour and 6-hour models are nearly identical, consistent with a shared low-loss region of weight space, while the 1-hour model has moved to a different region. Along the straight line between the 3-hour and 6-hour weights, a single number λ selects a forecaster whose effective lead time tracks λ. We verify this at the two interior horizons that hourly ERA5 offers on this segment: at 0.25°, plain linear weight interpolation (LWI) at 4 and 5 hours beats every grid-space output-interpolation baseline on near-surface variables, with up to about $2\times$ less error energy at mesoscale wavelengths. Against autoregressive chains the trade-off is the expected one: the chain has lower geopotential point error but damps the spectrum, while LWI preserves the observed energy spectrum at half to a third of the chain's inference cost. Spherical-interpolation and Fisher-weighted variants give no improvement. Connectivity is local: it holds on the 3–6 h segment, decays as the horizon gap grows, and fails under extrapolation. The effect reproduces on two further, architecturally different models (Pangu-Weather and GraphCast) and is stable across training seeds. We present this as an empirical characterization of when weight-space interpolation across lead times works and where it fails; its use as forcing for coupled systems or in assimilation is the motivation, not a demonstrated result.

1 INTRODUCTION

Machine-learning weather models now produce global 0.25° forecasts in seconds and match or surpass operational numerical weather prediction (NWP) on many headline scores (Rasp et al., 2024; Ben Bouallègue et al., 2024). Aurora (Bodnar et al., 2025), AIFS (Lang et al., 2024), GraphCast (Lam et al., 2023), FourCastNet (Pathak et al., 2022; Bonev et al., 2025), Pangu-Weather (Bi et al., 2023), FuXi (Chen et al., 2023), ClimaX (Nguyen et al., 2023) and Stormer (Nguyen et al., 2024) differ in architecture but share one operational rigidity: each is trained for one, or a few, discrete lead times Δt. A 6-hour model outputs 6-hour forecasts. Reaching a horizon in between requires retraining, autoregressive chaining of shorter-step models (Bi et al., 2023), or interpolation of the outputs on the grid. Retraining is expensive. Chaining compounds the spectral blurring that data-driven models already exhibit (Ben Bouallègue et al., 2024; Olivetti & Messori, 2024). Grid-space interpolation displaces features without regenerating their small-scale content. This is more than an inconvenience: ocean, sea-ice, hydrology and air-quality models, and data-assimilation cycles that run below the 6-hour cadence, need atmospheric forcing at time steps unrelated to 6 hours, and they respond to the variance of that forcing, not only to its mean.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This paper asks a simple question: after fine-tuning one model to two neighboring lead times, how do the two solutions sit in weight space? If they share a low-loss basin, a predictor for a horizon between them should be obtainable by interpolating weights rather than by training, as the model-soup and mode-connectivity literature (Garipov et al., 2018; Draxler et al., 2018; Izmailov et al., 2018; Wortsman et al., 2022) would suggest for fine-tunes of a common initialization. Here the two "tasks" are forecast horizons of one physical system, so an interpolation that works also says something about the temporal structure of the learned model.

We instantiate the question on the public Aurora 6-hour checkpoint, from which we fine-tune 3-hour and 1-hour adaptations with an identical recipe (Section 3). Their representations reveal an asymmetry: the 3h and 6h models are close (mean diagonal CKA ≥ 0.97 on four of seven stages and ≥ 0.55 on the rest; orthogonal Procrustes agrees), whereas the 1h model has reorganized (CKA ≤ 0.06 at the four stages where the 3h–6h pair saturates and ≤ 0.18 elsewhere). We therefore interpolate on the 3h–6h segment, and then test whether the result is specific to Aurora by repeating the recipe on two other architectures. Our findings are:

- **A low-loss valley in the lead-time direction.** At a fixed interior target horizon $T^\star$, the loss along the straight line between the 3h and 6h weights is U-shaped: at $T^\star = 4$ h its minimum, at $\lambda^\star \approx 0.30$, lies 45% below the better endpoint, and $\lambda^\star$ shifts monotonically with $T^\star$ (0.60 at 5 h). The valley location and depth are the same for three independent fine-tuning seeds (Section 4).
- **Three architectures, with controls.** The valley reproduces on Pangu-Weather, a 3D transformer with no lead-time input at all, and on GraphCast, a graph network, with minima 37–48% below the better endpoint. Interpolating Pangu-Weather's officially released, *independently trained* 3h and 6h models instead crosses a loss barrier at every interior λ (peak error $27\times$ the better endpoint). A shared fine-tuning basin is what makes the segment interpolable, not adjacency of horizons alone (Section 4.1).
- **Usable 4h and 5h forecasts given two endpoints.** Evaluated at every 6-hourly initialization of a held-out year (2022, $N = 1460$), LWI beats every grid-space interpolant on near-surface variables and trades geopotential point error against spectral fidelity relative to autoregressive chains, at half to a third of their inference cost. Natively fine-tuned 4h and 5h models, trained as an upper reference, lead LWI by 7–33% in RMSE at the price of one additional fine-tuning run per horizon (Section 5).
- **Spectral fidelity.** LWI preserves the ERA5 angular power spectrum; its error-field energy at meso-α scales is up to $2\times$ lower than that of grid-space baselines (Section 6).
- **Locality.** Extrapolation in λ fails, connectivity decays across the 1h–3h gap, and iterating an interpolant beyond its segment is stable but trails autoregressive chains (Sections 7 and 9).

The scope is deliberately the segment between two adjacent fine-tuned checkpoints, and the contribution is empirical: a characterization of when interpolation across lead times works, on which architectures, at what cost in skill relative to native models and chains, and where it fails. We do not demonstrate downstream use and do not claim a replacement for a dynamical core.

2 Related work

Data-driven weather models. High-resolution data-driven forecasting is dominated by transformer (Bi et al., 2023; Nguyen et al., 2023; Bodnar et al., 2025; Nguyen et al., 2024), neural-operator (Pathak et al., 2022; Bonev et al., 2025), graph-network (Keisler, 2022; Lam et al., 2023; Lang et al., 2024) and diffusion (Price et al., 2025) systems trained on ERA5 (Hersbach et al., 2020) and benchmarked through WeatherBench 2 (Rasp et al., 2020; 2024). Their forecasts are produced at a small set of lead times fixed by data sampling and rollout design. FuXi (Chen et al., 2023) is a 6-hourly cascade, and reaching hourly output required a new model, FuXi-2.0 (Zhong et al., 2024); Pangu-Weather (Bi et al., 2023) trains independent models for several lead times and combines them greedily, so chained forecasts inherit the spectral mismatch of each component. We never train a model for the intermediate horizon; we synthesize it from two existing ones in weight space.

Weight interpolation, mode connectivity and task arithmetic. Distinct solutions can be connected by low-loss curves (Garipov et al., 2018; Draxler et al., 2018), averaging fine-tuned weights can match

or exceed the individual checkpoints (Izmailov et al., 2018; Wortsman et al., 2022), and fine-tunes of a shared initialization tend to lie in a common basin (Frankle et al., 2020). In the task-arithmetic view (Ilharco et al., 2023; Ortiz-Jimenez et al., 2023), the difference $\Delta\mathcal{W} = \mathcal{W}_{6\text{h}} - \mathcal{W}_{3\text{h}}$ is a task vector and $\mathcal{W}_{3\text{h}} + \lambda\Delta\mathcal{W}$ a scaled edit; Appendix E.2 shows that $\Delta\mathcal{W}$ is dominated by a handful of singular directions per stage, the regime in which a single scalar edit is well posed. Two aspects distinguish our setting. First, the goal is not better performance on the original task but the synthesis of a *new* task, forecasting at an intermediate horizon. Second, the axis is physical: neighboring points correspond to neighboring horizons of one dynamical system. The closest precedents are Kangaslahti & Alvarez-Melis (2025), who interpolate low-rank adapters of a frozen language model along stylistic attribute axes and obtain smooth, predictable intermediates, and Duncan et al. (2024), who interpolate a pretrained and a fine-tuned atmospheric emulator and find that climate biases vary approximately linearly with λ. In both cases the axis is model identity or style, so interior points are compromises between the endpoints. In our case interior points are distinct, well-posed tasks, and the loss has an interior *minimum*. Madaan et al. (2026) report that interpolating or extrapolating checkpoints toward an unobserved future data distribution does not beat simply using the latest checkpoint, even when consecutive checkpoints are kept close. The difference is target observability: their target lies outside the observed data; ours is an interior point bracketed by two anchors trained on observed data of the same system, which supplies exactly the assumption about the data-generating process that they identify as the missing ingredient. Our negative results agree with theirs: extrapolation outside $[0, 1]$ fails, and independently trained checkpoints do not interpolate.

Continuous-time and explicit-time forecasting. Neural ODEs and latent ODEs (Chen et al., 2018; Rubanova et al., 2019) parameterize time explicitly; NeuralGCM (Kochkov et al., 2024) couples a learned component to a dynamical core at coarser grids than direct 0.25° predictors. In data-driven weather models, lead-time embeddings (Sønderby et al., 2020; Espeholt et al., 2022) and randomized-interval objectives (Nguyen et al., 2024) condition a single network on the horizon during training. Our approach is complementary: we recover an implicit temporal continuum from two discrete checkpoints without changing how they were trained.

Why 6 h, 3 h and 1 h. Six hours is the canonical NWP cadence and the lead time of the released checkpoints; three hours is the coarsest sub-step that admits a direct chain (3h + 3h $\equiv$ 6h); one hour matches operational temporal resolution and gives the non-interpolation baseline, the chain $\mathcal{W}_{1\text{h}} \to \mathcal{W}_{3\text{h}}$. We use Centered Kernel Alignment (CKA; Kornblith et al., 2019) to characterize the fine-tuned checkpoints; in our regime high CKA coincides with interpolation success and low CKA with failure, and we use it as a coarse indicator, not a criterion.

3 SETUP

Models and fine-tuning. Our primary test bed is the public Aurora foundation model (Bodnar et al., 2025), which takes two atmospheric states $(t, t - \Delta t)$ as input and predicts the state at $t + \Delta t$. Starting from the released $\Delta t = 6$ h checkpoint, we fine-tune a 3-hour predictor $\mathcal{W}_{3\text{h}}$ and a 1-hour predictor $\mathcal{W}_{1\text{h}}$ on 6-hourly ERA5 (Hersbach et al., 2020) (initialization times 00, 06, 12 and 18 UTC). *The two adaptations share an identical optimization recipe* (learning-rate schedule, number of steps, optimizer, batch composition) and differ only in the input spacing and the forecast horizon of a single step: the 3h model takes inputs at $(t, t - 3\text{h})$ and the 1h model at $(t, t - 1\text{h})$, so that the two-state stencil is symmetric with respect to the target. The number of fine-tuning steps is about 2.4% of the original pretraining effort. To avoid loss-imbalance pathologies across the heterogeneous variable set, the per-variable loss weights are initialized to the inverse of the average per-variable loss measured on a validation pass before fine-tuning; the surface–atmosphere balance factor of 0.25 is preserved (Appendix A). The 6h model is the original checkpoint. With the same recipe we additionally train native 2h, 4h and 5h adaptations, which serve as upper references (Section 5) and are not used by our method. To test architecture dependence we repeat the fine-tune-then-interpolate recipe on **Pangu-Weather** (Bi et al., 2023), a 3D Swin-style transformer that takes a single input state and has no lead-time input at all (each released lead time is a separate, independently trained network; we operate directly on the learnable tensors of the released ONNX graphs), and on **GraphCast** (Lam et al., 2023) in its operational configuration (0.25°, 13 pressure levels), a graph network. For Pangu-Weather a 3h anchor is fine-tuned *from* the official 6h weights; for GraphCast both anchors are fine-tuned from the public checkpoint under one pipeline. Both use a single pass over seven years

of ERA5, a deliberately light budget: these are proofs of concept, not state-of-the-art adaptations (Appendix A.4).[1]

Data and verification. Fine-tuning of the Aurora adaptations uses ERA5 2008–2020 with 2021 for validation; 2022 is never seen by any model and is the test year (hourly $0.25°$ ERA5). Two samples of it are used: *every* 6-hourly initialization ($N = 1460$) for the headline comparison (Table 2), the native baselines, the rollouts and the single-step tables of the other architectures, stratified by season in Appendix C; and a random subset of 64 initializations for the complete baseline set with ACC (Appendix C), the spectra (Section 6) and all λ-sweeps. Two evaluation protocols are used, and they are never mixed. In the *forecast protocol* of every table, a model evaluated at horizon T receives the input pair $(t - T, t)$ and the lead-time conditioning T. In the *sweep protocol* of Section 4, the input pair is pinned to the 3h anchor's stencil $(t - 3\text{h}, t)$ for every λ, endpoints included, while conditioning and target offset are set to $T^\star$, so that the whole path is scored on one loss surface. The same weights therefore take different absolute values under the two protocols (Appendix E.6); "LWI" in the abstract and in the claims refers to the forecast protocol. We report latitude-weighted RMSE and anomaly correlation (ACC, against the WeatherBench-2 climatology (Rasp et al., 2024)); definitions are in Appendix A.2. Because point-wise metrics are blind to spectral bias we additionally compute the angular power spectrum of forecast and error fields with `pyshtools` (Wieczorek & Meschede, 2018), integrated over a synoptic band ($10 \le \ell < 60$, wavelengths of roughly 700–4000 km) and a meso-α band ($60 \le \ell < 200$, roughly 200–700 km), together with the spectral slope over $\ell \in [10, 100]$. Method-vs-method differences are assessed with a paired bootstrap over initializations ($B = 10{,}000$ resamples; Appendix D); "significant" below means the 95% interval on ΔRMSE excludes zero. Z500 RMSE is reported in m^2/s^2 throughout.

Sanity checks. Both $\mathcal{W}_{3\text{h}}$ and $\mathcal{W}_{1\text{h}}$ reach plausible single-step skill (Appendix B), and rolling $\mathcal{W}_{3\text{h}}$ twice reproduces the 6-hour skill of the original checkpoint within noise, so the adaptation has not damaged the 6h skill encoded in the weights.

Representational geometry of $\mathcal{W}_{1\text{h}}, \mathcal{W}_{3\text{h}}, \mathcal{W}_{6\text{h}}$. Layer-wise CKA between $\mathcal{W}_{3\text{h}}$ and $\mathcal{W}_{6\text{h}}$ saturates (≥ 0.97) on four of the seven architectural stages and stays at 0.55–0.74 on the three mid-depth blocks (Table 14, Appendix E.4), whereas every entry pairing $\mathcal{W}_{1\text{h}}$ with either of the others collapses to ≤ 0.06 at those four stages and stays ≤ 0.18 at the mid-depth blocks. The 3h and 6h checkpoints therefore share representations, while the 1h adaptation has reorganized onto a different basin despite the matched budget. Two further diagnostics in Appendix E agree: weight deltas are small and dominated by a few singular directions per stage, with the 1h–6h delta consistently larger than 3h–6h, and the effective spatial reach of attention grows monotonically with lead time ($\mathcal{W}_{1\text{h}} \sim 350$ km, $\mathcal{W}_{3\text{h}} \sim 390$ km, $\mathcal{W}_{6\text{h}} \sim 557$ km), as expected if longer horizons are dominated by synoptic-scale features. This is what motivates interpolating between 3h and 6h and not across 1h.

4 Low-loss connectivity between $\mathcal{W}_{3\text{h}}$ and $\mathcal{W}_{6\text{h}}$

Linear weight interpolation. We synthesize an intermediate-horizon predictor as

$$\mathcal{W}_\lambda \;=\; (1-\lambda)\,\mathcal{W}_{3\text{h}} + \lambda\,\mathcal{W}_{6\text{h}}, \qquad \lambda \in [0,1], \tag{1}$$

with all parameters interpolated, including the affine parameters of the normalization layers (LayerNorm keeps no running statistics, so there is nothing to omit). The nominal map $T(\lambda) = 3 + 3\lambda$ h gives $\lambda_{4\text{h}} = 1/3$ and $\lambda_{5\text{h}} = 2/3$, which we use in Table 2; the sweeps below locate the actual optimum. $\mathcal{W}_\lambda$ is a single network of the original size, so a forecast costs exactly one forward pass.

A low-loss valley along the segment. We measure $L(\mathcal{W}_\lambda)$ at a *fixed* interior target horizon $T^\star$, with the lead-time conditioning and the target offset pinned to $T^\star$ and the input stencil pinned to $(t - 3\text{h}, t)$ for every λ, so that all points on the path, endpoints included, are scored on one loss surface (the endpoints are wrong-horizon evaluations of the native checkpoints under a fixed stencil; absolute values therefore differ from the forecast protocol of Table 2, see Section 3). The convex-combination bound $L(\mathcal{W}_\lambda) \le (1-\lambda)L(\mathcal{W}_{3\text{h}}) + \lambda L(\mathcal{W}_{6\text{h}}) + \frac{\kappa}{2}\lambda(1-\lambda)\|\Delta\mathcal{W}\|^2$, the generic

[1] We also attempted AIFS (Lang et al., 2024); end-to-end fine-tuning requires 18 additional input variables that our data pipeline does not provide, so it is not included.

smoothness inequality with curvature constant κ, is only motivating context: on every segment we measured, the realized loss lies strictly *below* the chord $L_{\text{lin}}(\lambda)$ at every interior λ, so the bound holds with $\kappa = 0$ and the substantive finding is the valley itself. On Aurora at $T^\star = 4\,\text{h}$ (Z500, full 21-point sweep, Appendix E.6) the endpoints are $L(\mathcal{W}_{3\text{h}}) = 44.94$ and $L(\mathcal{W}_{6\text{h}}) = 102.74$, while $\min_\lambda L(\mathcal{W}_\lambda) = 24.75$ at $\lambda^\star \approx 0.30$: the minimum sits 45% below the better wrong-horizon endpoint. At $T^\star = 5\,\text{h}$ the minimum is 32.92 at $\lambda^\star \approx 0.60$ against endpoints 87.66 and 70.51, a 53% gap. $\lambda^\star$ shifts monotonically with $T^\star$. Spherical interpolation, Fisher-weighted interpolation and interpolation of inter-block weights only match but do not improve plain LWI (Appendix E); the endpoints have near-unit cosine similarity, so SLERP reduces to LERP, and freezing any weight subset at one endpoint degrades the result on a small sample (Appendix E.8), so there is nothing for a reweighting to fix.

Seed control. To separate the geometry of the lead-time adaptation from the luck of one trajectory, we repeated the 3h fine-tune with three independent seeds and recomputed every diagnostic. The results are near-identical: the CKA saturation pattern (four to five of seven stages ≥ 0.97), the valley location ($\lambda^\star = 0.30$ for every seed at the 0.05 sweep resolution) and its depth (24.5–25.1 at the minimum, 44–46% below the better endpoint) all reproduce (Appendix E.7).

4.1 Beyond Aurora: Pangu-Weather and GraphCast

Aurora is a favorable case in two respects: it is conditioned on the lead time explicitly, and it is comparatively smooth. Pangu-Weather is the opposite case on both counts, and it ships something Aurora does not: officially released 3h and 6h models that were trained *independently*. This allows a two-sided experiment (Figure 1, Table 1).

Negative control. Linearly interpolating the official, independently trained Pangu-Weather 3h and 6h checkpoints crosses a loss barrier at every interior λ (same fixed-$T^\star$ protocol, $T^\star = 4\,\text{h}$): Z500 RMSE rises from the endpoint values 57.3 and 98.2 to a peak of 1563 at $\lambda \approx 0.5$, $27\times$ the better endpoint, and T2m at $\lambda = 1/3$ degrades to about $7\,\text{K}$ against endpoint values of 0.50 and $0.57\,\text{K}$, with visible structural artifacts (Appendix F). The endpoint separation $\|\Delta\mathcal{W}\|_F = 4767$, against 46 for the Aurora pair and 11 for the fine-tuned Pangu pair below, independently indicates that the two official checkpoints do not share a basin. Adjacent horizons alone are not sufficient.

Positive control. Fine-tuning a Pangu-Weather 3h model *from* the official 6h weights and repeating the sweep reproduces the valley: at $T^\star = 4\,\text{h}$ the minimum is 30.58 at $\lambda^\star \approx 0.50$, 48% below the better endpoint (59.21/86.21); at $T^\star = 5\,\text{h}$ it is 28.93 at $\lambda^\star \approx 0.80$, 38% below (99.01/46.79); $\varepsilon(\lambda) = L(\mathcal{W}_\lambda) - L_{\text{lin}}(\lambda) < 0$ at every interior point of both sweeps. Because Pangu-Weather has no lead-time input, every model on the path is scored on literally the same task, and the valley cannot be an artifact of interpolating conditioning parameters: the interpolable structure lives in the bulk weights. CKA between the fine-tuned 3h model and the official 6h checkpoint saturates (≥ 0.976) on seven of eight stages with its minimum at mid-depth (0.929), the same signature as on Aurora. Even at the uncalibrated nominal $\lambda = 1/3$ and $2/3$, the blends give usable 4h and 5h forecasts (Z500 34.6 and 34.8, against wrong-horizon endpoints 59.2/86.2 at 4 h and 99.0/46.8 at 5 h).

GraphCast. The same recipe on GraphCast, a graph network, gives minima 37% ($T^\star = 4\,\text{h}$, $\lambda^\star \approx 0.30$) and 38% ($T^\star = 5\,\text{h}$, $\lambda^\star \approx 0.60$) below the better endpoint, again with $\varepsilon(\lambda) < 0$ everywhere, and again with CKA saturating in the encoder (1.000) and dipping mid-network (processor 0.940, decoder 0.930).

What transfers and what does not. Three things transfer: the interior minimum, its monotone shift with $T^\star$, and the mid-depth concentration of the representational change. The calibration of λ to horizon does not: Aurora's and GraphCast's optima sit at 0.30 and 0.60, Pangu-Weather's at 0.50 and 0.80. λ measures displacement in weight space, not lead time, and in Pangu-Weather all horizon behavior must be carried by the weights; the map is calibrated once per pair with such a sweep, at the cost of a few dozen forward passes.

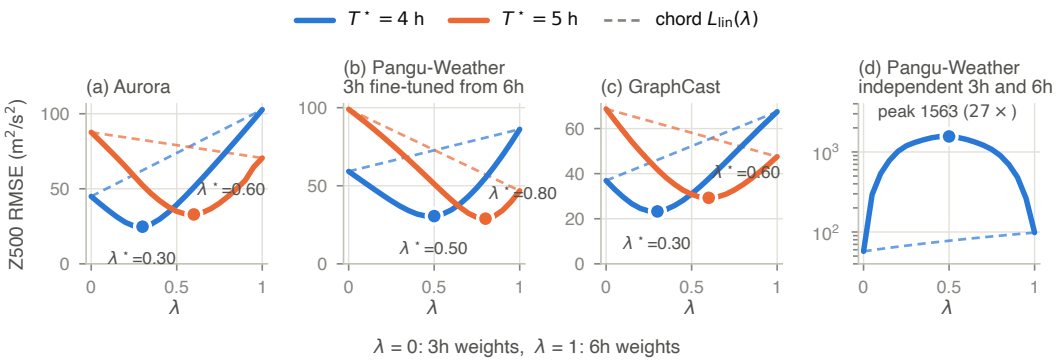


Figure 1: **The low-loss valley across three architectures, and its absence without a shared basin.** Z500 RMSE (m^2/s^2) of the interpolated weights $\mathcal{W}_\lambda = (1-\lambda)\mathcal{W}_{3h} + \lambda\mathcal{W}_{6h}$ evaluated at a fixed target horizon $T^\star = 4$ h and 5 h (solid), against the convex-combination chord $L_{\text{lin}}(\lambda)$ (dashed). (a) Aurora, (b) Pangu-Weather with the 3h anchor fine-tuned from the official 6h weights, (c) GraphCast: the realized loss lies below the chord at every interior λ, the minimum moves to larger λ for the longer target horizon, and the calibration of λ to horizon is pair-specific. (d) The officially released, independently trained Pangu-Weather 3h and 6h checkpoints (logarithmic axis): the loss rises above both endpoints at every interior λ, peaking at 27× the better one. Full sweeps are tabulated in Appendices E.6 and F.

Table 1: Cross-architecture summary of the fixed-$T^\star$ sweeps. Depth is the relative gap between the interior minimum and the better endpoint at that $T^\star$; the CKA column summarizes the stage-wise similarity between the two anchors (min. = lowest stage). $\lambda^\star$ is reported at the 0.05 resolution of the sweeps. The independently trained Pangu-Weather pair has no interior minimum at all.

Model	Architecture	CKA(3h, 6h)	$\lambda^\star$(4h)	depth	$\lambda^\star$(5h)	depth
Aurora	Swin transformer, lead-time input	4/7 stages ≥ 0.97; min. 0.55	0.30	45%	0.60	53%
Pangu-Weather (fine-tuned 3h)	3D Swin transformer, no lead-time input	7/8 stages ≥ 0.976; min. 0.93	0.50	48%	0.80	38%
GraphCast (operational)	graph network	encoder 1.00; processor 0.94; decoder 0.93	0.30	37%	0.60	38%
Pangu-Weather (official 3h, 6h)	as above; trained independently	—	loss barrier at every interior λ; peak 27× the better endpoint			

5 Forecast skill at the native sub-steps (4 h and 5 h)

Table 2 compares LWI at the nominal $\lambda = 1/3$ and $2/3$ with the strongest grid-space interpolant (a cubic spline through the 3h and 6h forecasts; linear and quadratic interpolants, linear-tendency extrapolation and persistence are uniformly worse and are reported in Appendix C), with the autoregressive chain of the native fine-tuned models ($\mathcal{W}_{1h} \to \mathcal{W}_{3h}$ for 4 h, $\mathcal{W}_{1h} \to \mathcal{W}_{1h} \to \mathcal{W}_{3h}$ for 5 h), and with natively fine-tuned 4h and 5h models, on every 6-hourly initialization of 2022. Five observations.

(a) Near-surface variables. LWI beats the cubic spline, and with it every grid-space interpolant, on T2m, T850, U10 and V10 at both horizons; all eight paired-bootstrap intervals exclude zero (Appendix D). The margins are largest on the winds (0.06–0.10 m/s, i.e. 12–18%). **(b) Geopotential.** On Z500 the autoregressive chain has lower point-wise error ($\Delta\text{RMSE}_{\text{LWI}-\text{Chain}} = +4.8$ at 4 h and $+5.4$ at 5 h, significant), and the cubic spline edges LWI by 0.9 and 0.8. This is a spectrum-versus-RMSE trade-off: the chain accumulates little geopotential error over $1 + 3$ h but damps the synoptic spectrum, and the spline is a smoother; LWI preserves the ERA5 spectrum and matches or beats the chain in meso-α error energy on the surface variables (Section 6), at one-third of the chain's inference cost at 5 h. We therefore do not claim that LWI beats the chain on Z500. **(c) Native models.** The natively fine-tuned 4h and 5h models lead LWI by 30–33% on Z500, 11–28% on T2m and 7–10% on T850 and the winds, and lead the chain on every variable, though on Z500 only within one standard deviation. Each native horizon costs one additional fine-tuning run (Appendix A.4); LWI needs only

Table 2: RMSE (lower is better) at 4 h and 5 h on Aurora, mean ± std over every 6-hourly initialization of the held-out year 2022 ($N = 1460$; Z500 in m^2/s^2, T2m and T850 in K, winds in m/s). LWI uses the nominal $\lambda = 1/3$ (4 h) and $2/3$ (5 h); *Chain* is $\mathcal{W}_{1h} \to \mathcal{W}_{3h}$ at 4 h and $\mathcal{W}_{1h} \to \mathcal{W}_{1h} \to \mathcal{W}_{3h}$ at 5 h; the cubic spline interpolates the 3h and 6h forecasts on the grid. **Bold**: best among the three methods that need no additional training (ties within the paired-bootstrap interval are both bold). The native rows are separately fine-tuned models and require one extra training run per horizon; they are shown as an upper reference. Paired-bootstrap intervals for every comparison are in Table 9.

Lead (h)	Method	Z500	T2m	T850	U10	V10
4	Cubic spline (grid)	25.11 ± 6.89	0.61 ± 0.10	0.39 ± 0.08	0.50 ± 0.11	0.54 ± 0.11
	LWI (ours), $\lambda = 1/3$	25.97 ± 4.42	$\mathbf{0.59 \pm 0.07}$	$\mathbf{0.37 \pm 0.08}$	$\mathbf{0.44 \pm 0.10}$	$\mathbf{0.44 \pm 0.11}$
	Chain 1h + 3h	$\mathbf{21.16 \pm 2.64}$	0.63 ± 0.02	0.38 ± 0.03	0.46 ± 0.05	0.46 ± 0.05
	Native 4h (reference)	20.01 ± 5.82	0.46 ± 0.02	0.34 ± 0.08	0.40 ± 0.10	0.40 ± 0.11
5	Cubic spline (grid)	25.72 ± 5.00	0.60 ± 0.04	0.40 ± 0.07	0.52 ± 0.08	0.56 ± 0.08
	LWI (ours), $\lambda = 2/3$	26.55 ± 4.59	$\mathbf{0.52 \pm 0.03}$	$\mathbf{0.38 \pm 0.07}$	$\mathbf{0.44 \pm 0.09}$	$\mathbf{0.46 \pm 0.09}$
	Chain 1h + 1h + 3h	$\mathbf{21.12 \pm 2.04}$	0.53 ± 0.04	0.39 ± 0.02	0.46 ± 0.04	$\mathbf{0.46 \pm 0.05}$
	Native 5h (reference)	19.93 ± 5.77	0.47 ± 0.03	0.35 ± 0.07	0.41 ± 0.10	0.42 ± 0.11

the two endpoints. **(d) Wrong-horizon collapse.** Asking $\mathcal{W}_{3h}$ itself to output a 4 h or 5 h forecast gives Z500 RMSE of 48.7 and 97.0 (LWI: 23.7 and 23.9 on the same 64-initialization subset under the forecast protocol, Appendix C; the sweep endpoints of Section 4 differ because of the pinned stencil), and $\mathcal{W}_{6h}$ degrades comparably in the opposite direction, so the interpolants are not "almost the same model" as an endpoint: the path traces distinct forecasters. **(e) Seasons and ACC.** The ranking is the same in every season (Appendix C): the chain leads on Z500 in all four, LWI leads or ties on T2m in all four with the largest margin in JJA (0.59 vs. 0.67 K). ACC is saturated (> 0.999 on Z500 for every viable method; Appendix C), so RMSE is the discriminating metric.

6 Spectral fidelity

Point-wise metrics undersell what is, in our view, the most physically interesting property of LWI: it preserves spectra. We compute the angular power spectrum of the error field $f - o$ on Z500 and the surface variables and integrate it over the synoptic and meso-α bands (Figure 2).

4 h and 5 h. LWI has lower error-field energy than every grid-space baseline in every variable and band. The largest gaps are in the meso-α band of the surface winds: LWI's V10 error energy is $2.1\times$ smaller than the cubic spline's at both horizons ($2.06\times$ at 4 h, $2.12\times$ at 5 h), and its U10 error energy at 5 h $1.8\times$ smaller. The mechanism is direct. A spline in observation space displaces small-scale features without regenerating their high-frequency content, whereas LWI inherits the spectral characteristics of $\mathcal{W}_{3h}$ and $\mathcal{W}_{6h}$, both trained against the high-frequency content of ERA5. Even against the autoregressive chain, LWI matches or beats it in meso-α energy on the surface variables despite the chain's Z500 advantage, a trade-off familiar from the literature (Rasp et al., 2024; Bonev et al., 2025).

Spectral skill and slope. The synoptic-to-planetary spectral skill of LWI lies near unity. Spectral slopes do not discriminate methods in our regime: across LWI, the chain and every grid-space interpolant the Z500 slope clusters at $\hat{\beta} \approx -3.44$ to -3.47 in the synoptic band and ≈ -2.12 in meso-α (not tabulated), comparable to the native endpoints; the discriminating quantity is the per-band error energy, which separates methods by 1.4–$2.1\times$.

7 Iterating an interpolant beyond its segment

The claims above concern a single forecast step. Rolling $\mathcal{W}_{1/3}$ (4 h) and $\mathcal{W}_{2/3}$ (5 h) autoregressively on their own outputs for two to four steps, on every initialization of 2022 (Appendix G), gives stable and physically coherent forecasts: Z500 RMSE grows from 32.1 at +8 h to 52.0 at +16 h for the 4h interpolant and from 35.5 at +10 h to 51.8 at +20 h for the 5h one, with T2m between 0.65 and 0.84 K. They are clearly worse than the native 6h model rolled to comparable horizons (Z500 25.2 at +12 h on the 64-initialization subset), by a factor of about 1.6 on Z500 and 1.3 on T2m. The

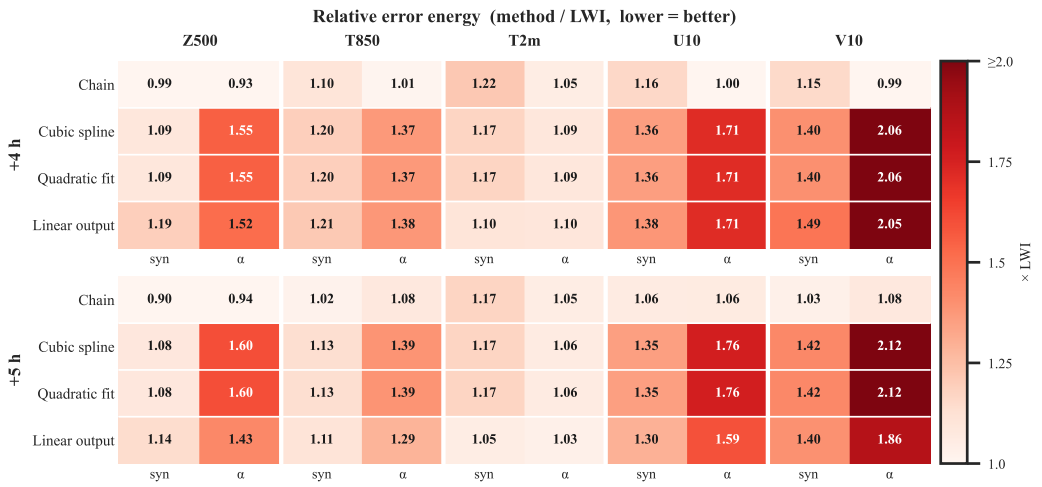


Figure 2: **Spectral error energy of every baseline, normalized by LWI** (lower is better; values below 1 mean the baseline beats LWI). Each cell is the ratio of band-integrated error power $\sum_{\ell \in \mathcal{B}} (2\ell + 1)\, C_\ell^{\mathrm{err}}$ of a baseline to that of LWI at $+4$ h (top) and $+5$ h (bottom) on the 64-initialization subset of 2022. Columns group the five headline variables; within each, the synoptic band (*syn*, $10 \le \ell < 60$, roughly 700–4000 km) and the meso-α band (mα, $60 \le \ell < 200$, roughly 200–700 km). Rows compare the autoregressive chain and three output-space interpolants against LWI. LWI matches or beats every output-space baseline in every variable and band, with the largest gains in the meso-α band of the wind components (ratios up to $2.1\times$). The chain is the only competitor that edges below 1, on the smooth fields Z500 and T850 and marginally on V10 meso-α at 4 h (0.99); elsewhere it stays at 1.0–$1.2\times$ LWI. Color scale clipped at 2.0.

interpolants were never trained to be iterated (the endpoints use a two-step rollout loss at their own cadence), and a rollout-aware fine-tune of the endpoints is the natural next step; an interpolant is also a single fixed operator on its own output, unlike a chain of models never trained to compose, but that property does not offset the gap we measure. We report the rollouts as a limitation, not a use case.

8 Discussion

Cost accounting. Given two fine-tuned endpoint checkpoints, every horizon between them is available at zero additional training and at the inference cost of one forward pass. The method is not training-free in absolute terms: if only a public 6h checkpoint exists, a 3h endpoint must be fine-tuned once, at about 2.4% of pretraining compute.

Potential downstream use, not demonstrated here. Coupled and impact models (ocean, sea-ice, hydrology, air quality) run on time steps unrelated to 6 h and respond to the variance of the forcing; the spectral results of Section 6 are the property that would make interpolated forcing attractive where MSE-trained high-cadence chains blur, and a differentiable, spectrally faithful $\mathcal{W}_\lambda$ is a candidate forward operator for sub-6h assimilation cycles. Whether the point-error gap to native models and to chains (Sections 5 and 7) is acceptable for any of these systems is untested: establishing it requires coupled or assimilation experiments that this paper does not contain. These uses are our motivation, not a result.

What the evidence says about mechanism, and what it does not. CKA, the weight-delta geometry and the Pangu-Weather negative control are consistent with one reading: the two adaptations share a basin and the segment between them parameterizes lead time. They are correlates and a control, not a causal account; we have not identified which weights carry the horizon. That models sharing an architecture and a task but differing in one scalar physical condition lie, locally, on a low-dimensional manifold in weight space is suggested by three architectures and three seeds, not established; other dials such as resolution or forcing scenario are future work. Probabilistic forecasters (Price et al., 2025; Bonev et al., 2025) are outside our scope: interpolation there acts on a denoiser and must be

judged on the induced distribution (CRPS, spread–skill) rather than on RMSE and spectra of a single field.

9 LIMITATIONS

Extrapolation in λ does not work. Extending the segment to 1 h ($\lambda = -2/3$) or 8 h ($\lambda = 5/3$) by extrapolation fails: the resulting models produce blurred, dynamically inconsistent forecasts. $\mathcal{W}_\lambda$ stays in the 3h–6h basin only on $[0, 1]$.

Connectivity decays with the lead-time gap. Interpolating between $\mathcal{W}_{1\mathrm{h}}$ and $\mathcal{W}_{3\mathrm{h}}$ to reach 2 h gives reasonable but clearly inferior forecasts: the best of four input-spacing and conditioning configurations reaches Z500 RMSE 20.7, against 12.1 for the 1h + 1h chain and 11.7 for a natively fine-tuned 2h model (Appendix H). The $1.7\times$ gap mirrors the CKA collapse between $\mathcal{W}_{1\mathrm{h}}$ and the other two checkpoints. We did not study 1h–6h interpolation; by the same indicator the gap would be larger. Low-loss connectivity is a local property of the lead-time axis: it holds when the endpoints share a basin, decays with their distance in weight space, and, as the Pangu-Weather negative control shows, is absent between independently trained checkpoints.

Two adapted checkpoints are required, and λ must be calibrated per pair. We only tested endpoints that share a pretraining initialization, in line with the model-soup regime (Wortsman et al., 2022). The map from λ to effective horizon is pair-specific (Table 1) and should be calibrated with a short sweep before use.

The continuous-horizon reading is verified at two points. Hourly ERA5 provides only two interior targets on the 3–6 h segment, 4 h and 5 h. The sweeps show a smooth loss in λ and a monotone shift of the optimum between these two targets, but we have no verification at any other horizon, and a segment with more interior hours (for example 6–12 h) is the natural test of the continuous reading.

Iterated rollouts trail native chains. Section 7 quantifies the gap; a rollout-aware fine-tuning recipe for the endpoints is future work.

Light-budget anchors on Pangu-Weather and GraphCast. The 3h anchors on the two additional architectures are proofs of concept trained with a deliberately small budget: the fine-tuned Pangu-Weather 3h model has Z500 RMSE 22.2 at its native lead against 13.1 for the official 3h model. The valley appears regardless, which is the point; the absolute skill of the interpolants on those two models is not competitive with well-tuned native checkpoints.

Statistical scope. All comparisons rest on a single held-out year (2022), at full temporal coverage for the headline tables and on a 64-initialization subset for the diagnostics; additional years and variables, and extreme-event case studies, which a uniform sample cannot characterize, are left to future work.

10 CONCLUSION

Two adjacent lead-time adaptations of a weather model are connected by a low-loss segment in weight space, and linear interpolation along it yields forecasters at the intermediate horizons we can verify that are spectrally faithful, competitive with output-space and chain baselines at a single step, and free once the endpoints exist. The segment has an interior optimum whose position tracks the target horizon; it is the same for three training seeds and reproduces on Pangu-Weather and GraphCast, and it does not exist between independently trained checkpoints. The evidence is consistent with lead time being, locally, a low-dimensional axis in the post-fine-tuning weight space of these models; we have characterized where that holds and where it breaks, not why. The design implication is modest: fine-tune at a sparse set of physically motivated endpoints and recover the interior by interpolation, wherever two adjacent adaptations of a shared backbone share a basin.

AI USE STATEMENT

In this work, we used generative AI tools for implementing analysis and plotting scripts on top of our in-house training library, for preparing the standalone code release, and for editing and restructuring the manuscript, including the abstract and the appendices, from the authors' text, outlines and results; the authors revised every section. We also used them to propose hypotheses and experiments; those we tested performed worse than our own designs, and none is part of the final method. Additionally, we used generative AI tools for literature search, discussion of the methodology and arguments, and a mock review of the draft. The code release was tested end to end on a fresh machine following its README; the numbers of the main text and of the appendices are produced by generator scripts from the evaluation artifacts; the references were checked against their DOIs or arXiv records by the authors. We take responsibility for the final content of this work, including text, claims or artefacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All models are public: the Aurora checkpoint (Bodnar et al., 2025), the Pangu-Weather ONNX models (Bi et al., 2023) and the GraphCast operational checkpoint (Lam et al., 2023); all data are public ERA5 (Hersbach et al., 2020) and WeatherBench-2 climatologies (Rasp et al., 2024). The fine-tuning recipe (optimizer, schedule, loss weighting, data split, budget) is given in Appendix A, the two evaluation protocols and the metric definitions in Section 3 and Appendix A.2, and every number quoted in the main text is tabulated in the appendices, including the complete λ-sweeps of all architectures (Appendices E.6 and F). The interpolation itself is a single line, Equation (1), applied to every learnable tensor. A standalone code repository that reproduces the weight interpolation, the baselines (chains, grid-space interpolants, wrong-horizon evaluations), the fixed-target sweeps and the paired bootstrap is provided as supplementary material and will be released publicly.

ETHICS STATEMENT

This work uses only publicly available reanalysis data and publicly released model weights, under their respective licenses (Appendix A.5); the released Aurora, Pangu-Weather and GraphCast weights carry non-commercial licenses, and our fine-tuned anchors inherit them. No human subjects or personal data are involved. Better sub-6-hour atmospheric forcing would benefit downstream forecasting of hazards (floods, air quality, sea ice); we see no direct dual-use concern beyond those inherent to weather prediction, and we state the method's failure modes (extrapolation, non-adjacent horizons, iterated rollouts) explicitly so that it is not applied outside its validated regime.

REFERENCES

Zied Ben Bouallègue, Mariana C. A. Clare, Linus Magnusson, Estibaliz Gascón, Michael Maier-Gerber, Martin Janoŭšek, Mark Rodwell, Florian Pinault, Jesper S. Dramsch, Simon T. K. Lang, Baudouin Raoult, Florence Rabier, Matthieu Chevallier, Irina Sandu, Peter Dueben, Matthew Chantry, and Florian Pappenberger. The rise of data-driven weather forecasting: A first statistical assessment of machine learning-based weather forecasts in an operational-like context. *Bulletin of the American Meteorological Society*, 105(6):E864–E883, 2024. doi: 10.1175/BAMS-D-23-0162.1.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023. doi: 10.1038/s41586-023-06185-3.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the Earth system. *Nature*, 641(8065):1180–1187, 2025. doi: 10.1038/s41586-025-09005-y.

Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D. Collins, Michael S. Pritchard, and Alexander Keller. FourCastNet

3: A geometric approach to probabilistic machine-learning weather forecasting at scale, 2025. URL https://arxiv.org/abs/2507.12144.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6:190, 2023. doi: 10.1038/s41612-023-00512-1.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems 31*, 2018. URL https://dl.acm.org/doi/10.5555/3327757.3327764.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1309–1318. PMLR, 2018. URL https://proceedings.mlr.press/v80/draxler18a.html.

James P. C. Duncan, Elynn Wu, Jean-Christophe Golaz, Peter M. Caldwell, Oliver Watt-Meyer, Spencer K. Clark, Jeremy McGibbon, Gideon Dresdner, Karthik Kashinath, Boris Bonev, Michael S. Pritchard, and Christopher S. Bretherton. Application of the AI2 Climate Emulator to E3SMv2's global atmosphere model, with a focus on precipitation fidelity. *Journal of Geophysical Research: Machine Learning and Computation*, 1(3):e2024JH000136, 2024. doi: 10.1029/2024JH000136.

Lasse Espeholt, Shreya Agrawal, Casper Sønderby, Manoj Kumar, Jonathan Heek, Carla Bromberg, Cenk Gazen, Rob Carver, Marcin Andrychowicz, Jason Hickey, Aaron Bell, and Nal Kalchbrenner. Deep learning for twelve hour precipitation forecasts. *Nature Communications*, 13:5145, 2022. doi: 10.1038/s41467-022-32483-x.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3259–3269. PMLR, 2020. URL https://proceedings.mlr.press/v119/frankle20a.html.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems 31*, pp. 8803–8812, 2018. URL https://dl.acm.org/doi/10.5555/3327546.3327556.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Giorgio Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leopold Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Catalina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization, 2018. URL https://arxiv.org/abs/1803.05407.

Sara Kangaslahti and David Alvarez-Melis. Continuous language model interpolation yields dynamic and controllable text generation. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=xD9Nu2Wah4.

Ryan Keisler. Forecasting global weather with graph neural networks, 2022. URL https://arxiv.org/abs/2202.07575.

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, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024. doi: 10.1038/s41586-024-07744-y.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019. URL https://arxiv.org/abs/1905.00414.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, 2023. doi: 10.1126/science.adi2336.

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.

Divyam Madaan, Sumit Chopra, and Kyunghyun Cho. Temporal generalization: A reality check. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Wz0ILlbh9U.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. ClimaX: A foundation model for weather and climate. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25904–25938. PMLR, 2023. URL https://proceedings.mlr.press/v202/nguyen23a.html.

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Veerabhadra Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting, 2024. URL https://arxiv.org/abs/2312.03876.

Leonardo Olivetti and Gabriele Messori. Do data-driven models beat numerical models in forecasting weather extremes? A comparison of IFS HRES, Pangu-Weather, and GraphCast. *Geoscientific Model Development*, 17:7915–7962, 2024. doi: 10.5194/gmd-17-7915-2024.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems 36*, 2023. URL https://arxiv.org/abs/2305.12827.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Anima Anandkumar. FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators, 2022. URL https://arxiv.org/abs/2202.11214.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025. doi: 10.1038/s41586-024-08252-9.

Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Salva Mouatadid, and Nils Thuerey. WeatherBench: A benchmark dataset for data-driven weather forecasting, 2020. URL https://arxiv.org/abs/2002.00469.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallègue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems 32*, 2019. URL https://dl.acm.org/doi/10.5555/3454287.3454765.

Casper Kaae Sønderby, Lasse Espeholt, Jonathan Heek, Mostafa Dehghani, Avital Oliver, Tim Salimans, Shreya Agrawal, Jason Hickey, and Nal Kalchbrenner. MetNet: A neural weather model for precipitation forecasting, 2020. URL https://arxiv.org/abs/2003.12140.

Mark A. Wieczorek and Matthias Meschede. SHTools: Tools for working with spherical harmonics. *Geochemistry, Geophysics, Geosystems*, 19(8):2574–2592, 2018. doi: 10.1029/2018GC007529.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Xiaohui Zhong, Lei Chen, Xu Fan, Wenxu Qian, Jun Liu, and Hao Li. FuXi-2.0: Advancing machine learning weather forecasting model for practical applications, 2024. URL https://arxiv.org/abs/2409.07188.

A Training and evaluation details

A.1 Variables

Table 3: Variables used by the three models in the configurations considered here. A check mark indicates that the variable is an input and a prognostic output of the model; † marks a variable that is predicted but not used as input (GraphCast, operational configuration); × marks a variable the model does not use. All three models use the same 13 pressure levels. Pangu-Weather shares Aurora's variable set and static masks; GraphCast additionally carries vertical velocity, drops the soil-type mask, and takes solar and calendar forcings.

Acronym	Name	Type	Aurora	Pangu-Weather	GraphCast
Surface / single-level — prognostic					
2t	2 m temperature	surface	✓	✓	✓
10u	10 m zonal wind	surface	✓	✓	✓
10v	10 m meridional wind	surface	✓	✓	✓
msl	mean sea-level pressure	surface	✓	✓	✓
tp	total precipitation (6 h)	surface	×	×	†
Static / forcing (input only)					
lsm	land-sea mask	static	✓	✓	✓
slt	soil type	static	✓	✓	×
z_surf	surface geopotential (orography)	static	✓	✓	✓
tisr	TOA incident solar radiation	forcing	×	×	✓
—	year and day progress (sin/cos)	forcing	×	×	✓
Atmospheric (13 pressure levels) — prognostic					
z	geopotential	atmospheric	✓	✓	✓
u	zonal wind	atmospheric	✓	✓	✓
v	meridional wind	atmospheric	✓	✓	✓
t	temperature	atmospheric	✓	✓	✓
q	specific humidity	atmospheric	✓	✓	✓
w	vertical velocity	atmospheric	×	×	✓

Surface variables are defined at the Earth's surface or at fixed near-surface heights (2 m temperature, 10 m wind, etc.). Atmospheric variables are defined throughout the vertical column on pressure levels. Aurora uses the 13 pressure levels $\{50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000\}$ hPa. Lower-pressure levels correspond to higher altitudes. Pangu-Weather and the operational GraphCast checkpoint use the

same 13 levels, so the three models see the same vertical grid; the only prognostic variable present in one model and absent in the others is GraphCast's vertical velocity.

A.2 METRICS

We report latitude-weighted RMSE and anomaly correlation,

$$\mathrm{RMSE}(f,o)=\sqrt{\tfrac{1}{HW}\textstyle\sum_{i,j} w_i (f_{i,j}-o_{i,j})^2},\quad \mathrm{ACC}(f,o)=\frac{\sum_{i,j} w_i (f-\bar{o})(o-\bar{o})}{\sqrt{[\sum w_i (f-\bar{o})^2][\sum w_i (o-\bar{o})^2]}}, \tag{2}$$

where $f_{i,j}$ is the forecast at grid cell (i,j), $o_{i,j}$ the corresponding ERA5 value, $\bar{o}_{i,j}$ the per-cell, per-day-of-year WeatherBench-2 climatology, $H\times W$ the grid dimensions, and $w_i \propto \cos\phi_i$ latitude weights (ϕ_i the latitude of row i) renormalized to unit mean. Z500 is geopotential, so its RMSE is in $\mathrm{m^2/s^2}$ (divide by $g\approx 9.81\,\mathrm{m/s^2}$ for geopotential metres).

For the spectral diagnostics we compute the angular power spectrum $S_x(\ell)=\sum_m |a_{\ell m}|^2$ of a field x on the sphere with `pyshtools` (Wieczorek & Meschede, 2018) (an isotropic 2D-FFT fallback agrees up to a constant). From it we derive: the synoptic-to-planetary spectral skill $\mathrm{SpecSkill}=|\mathcal{L}|^{-1}\sum_{\ell\in\mathcal{L}} S_f(\ell)/S_o(\ell)$ with $\mathcal{L}=\{10,\dots,100\}$; the band-integrated energy of the error field, $\text{err-energy}(\mathcal{B})=\sum_{\ell\in\mathcal{B}} S_{f-o}(\ell)$, over the synoptic band $\mathcal{B}_{\text{syn}}=[10,60)$ and the meso-α band $\mathcal{B}_{\text{m}\alpha}=[60,200)$ (a meso-β band $[200,400)$ is also computed but not reported); the spectral slope $\hat{\beta}$, the least-squares slope of $\log S(\ell)$ against $\log\ell$ over a band, compared with the Nastrom–Gage reference $\hat{\beta}\approx -3$; and an effective spectral resolution $\ell^\star$, the wavenumber beyond which the per-ℓ spectral skill $S_f(\ell)/S_o(\ell)$ falls below a fixed threshold (0.9 and 0.5). Because the Z500 synoptic-band error energy has a high coefficient of variation across initializations, $\ell^\star$ is reported only on T2m, T850 and the surface winds.

A.3 LOSS FUNCTION

We use the standard surface/atmosphere-decomposed L1 loss

$$\mathcal{L}(\widehat{X}^t, X^t)=\frac{1}{V_S+V_A}\left[\gamma\,\mathcal{L}_{\text{surf}}+\mathcal{L}_{\text{atmos}}\right], \tag{3}$$

$$\mathcal{L}_{\text{surf}}=\sum_{k=1}^{V_S}\frac{w_k^S}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|\widehat{S}^t_{k,i,j}-S^t_{k,i,j}\right|, \tag{4}$$

$$\mathcal{L}_{\text{atmos}}=\sum_{k=1}^{V_A}\frac{1}{CHW}\sum_{c=1}^{C} w^A_{k,c}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|\widehat{A}^t_{k,c,i,j}-A^t_{k,c,i,j}\right|, \tag{5}$$

where $\widehat{X}^t=(\widehat{S}^t,\widehat{A}^t)$ and $X^t=(S^t,A^t)$ are the forecast and ground-truth surface (S) and atmospheric (A) variables; the spatial resolution is $H=720$ and $W=1440$, C is the number of pressure levels, $V_{S/A}$ count surface and atmospheric variables, $\gamma=1/4$ is the surface–atmosphere balance, and $w_k^{S/A}$ are the per-variable weights described in Section 3. The loss is computed across two rollout steps (the model forecasts, is fed its own forecast, forecasts again, and the gradient is taken on the sum), as in the original Aurora recipe.

A.4 DATA SPLITS AND FINE-TUNING RECIPES

For fine-tuning we use 8×80 GB GPUs. The year 2022 is used neither for training nor for validation and is the test year.

Native 2h, 4h and 5h adaptations. The native intermediate-horizon models (Section 5) use the same optimizer, schedule, loss and input-stencil convention as the 3h and 1h adaptations, with the input spacing and single-step horizon set to the target lead time. They are trained in a single pass over ERA5 2013–2020 (2021 for validation) on 8 GPUs, with the cosine schedule synchronized to 3-month pseudo-epochs, i.e. about 60% of the budget of the 3h and 1h adaptations.

Table 4: Training hyperparameters for the Aurora 3h and 1h adaptations.

Parameter	Value
Optimizer	AdamW
Learning rate	3×10^{-4}
LR schedule	Cosine, synchronized to 3-month pseudo-epochs
Rollout steps (loss)	2
Batch size	1 (with gradient checkpointing and CPU offload)
Train set	ERA5 2008–2020
Val set	ERA5 2021
Test set	ERA5 2022: every 6-hourly initialization ($N = 1460$) for the headline tables; a random 64-initialization subset for the complete baseline set, the spectra and the λ-sweeps
Number of fine-tuning steps	$\sim 2.4\%$ of the original pre-training compute (one pass over the training set on 8 GPUs)

Pangu-Weather. Pangu-Weather (Bi et al., 2023) is released as four ONNX graphs (1h, 3h, 6h and 24h), each an independently trained network with a single input state and no lead-time input. For the negative control we interpolate the learnable initializers of the official 3h and 6h graphs directly (convolution and linear weights and biases, and normalization affine parameters) over the name-aligned key set, leaving positional and static tensors untouched; there are no batch-normalization running statistics. For the positive control we fine-tune a 3h model from the official 6h weights with the recipe above (input spacing is not applicable, single-step target at 3 h) on a deliberately small budget: a single pass over seven years of 6-hourly ERA5 (2014–2020), with 2021 for validation, against the roughly 3000 V100-days per model that Bi et al. (2023) report for pretraining. The resulting 3h anchor is weaker than the official 3h model (Z500 RMSE 22.2 versus 13.1 at the native 3 h lead, Table 23); the valley of Section 4.1 appears regardless.

GraphCast. We use the operational GraphCast checkpoint (Lam et al., 2023) ($0.25°$, 13 pressure levels). Both anchors are fine-tuned from the public weights under the same pipeline and budget as the Pangu-Weather anchor (a single pass over ERA5 2014–2020, 2021 for validation): a 3h model with a 3 h single-step target, and a 6h model at the released cadence, so that the two endpoints share one data processing and one budget.

A.5 Licenses of existing assets

We rely on the following publicly available assets and use each under its original license.

- **ERA5 reanalysis** (Hersbach et al., 2020): Copernicus Climate Change Service (C3S) data, available from the Climate Data Store (CC-BY 4.0).
- **WeatherBench-2** (Rasp et al., 2024): Apache License 2.0 (climatology files and evaluation tooling).
- **Aurora** (Bodnar et al., 2025): model weights released under CC BY-NC-SA 4.0; we use the released $\Delta t = 6$ h checkpoint.
- **Pangu-Weather** (Bi et al., 2023): trained parameters released under CC BY-NC-SA 4.0 (ONNX graphs for 1h, 3h, 6h and 24h).
- **GraphCast** (Lam et al., 2023): model weights released under CC BY-NC-SA 4.0, code under Apache License 2.0; we use the operational checkpoint.
- **pyshtools** (Wieczorek & Meschede, 2018): BSD-3-Clause.

Our fine-tuned anchors inherit the non-commercial terms of the respective base weights. URLs and version pins are listed in the accompanying code archive.

B Native-model RMSE

Table 5: RMSE of the three Aurora anchor checkpoints, evaluated on ERA5 at $0.25°$ over the random 64-initialization subset of 2022 (mean $\pm$ std), obtained with the rollout evaluation script; single-step values differ slightly from Table 6, which comes from a separate evaluation run. Each row is one autoregressive rollout of the respective model: $\mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}}$ is rolled to $\{6, 12, 18, 24\}$ h; $\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}}$ to $\{3, 6\}$ h; $\mathcal{W}_{1\mathrm{h}}^{\mathrm{Aur}}$ to $\{1, 2, 3\}$ h. $\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}}$ rolled out twice (i.e. at 6 h) matches the native single-step $\mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}}$ within noise, confirming that the adaptation has not damaged the underlying 6 h skill.

Model	Rollout horizon (h)	RMSE ↓				
		T2m, K	MSL, Pa	Z500	T850	U850
$\mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}}$	6	0.45 ± 0.03	23.14 ± 4.49	18.81 ± 3.85	0.36 ± 0.06	0.65 ± 0.14
	12	0.51 ± 0.03	30.87 ± 1.16	25.21 ± 0.94	0.47 ± 0.02	0.88 ± 0.03
	18	0.59 ± 0.03	37.52 ± 3.47	31.95 ± 3.41	0.55 ± 0.04	1.04 ± 0.10
	24	0.65 ± 0.03	44.73 ± 2.17	39.88 ± 2.12	0.62 ± 0.02	1.20 ± 0.04
$\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}}$	3	0.45 ± 0.04	16.97 ± 0.82	12.91 ± 0.70	0.24 ± 0.01	0.40 ± 0.01
	6	0.48 ± 0.03	25.02 ± 5.42	18.91 ± 4.86	0.36 ± 0.06	0.67 ± 0.15
$\mathcal{W}_{1\mathrm{h}}^{\mathrm{Aur}}$	1	0.29 ± 0.02	12.05 ± 0.95	9.65 ± 0.81	0.18 ± 0.01	0.30 ± 0.01
	2	0.48 ± 0.04	15.90 ± 0.79	12.12 ± 0.73	0.23 ± 0.01	0.40 ± 0.01
	3	0.59 ± 0.06	19.65 ± 0.98	14.62 ± 0.89	0.28 ± 0.01	0.49 ± 0.01

C Full-year evaluation and the 64-initialization subset

The headline comparison of Table 2 uses every 6-hourly initialization of the held-out year 2022 ($N = 1460$). The complete set of baselines, together with ACC, was evaluated on a random subset of 64 initializations of the same year; those tables are reproduced here because they contain rows (persistence, linear-tendency extrapolation, linear and quadratic output interpolation, the wrong-horizon evaluations, and the native anchors at their own lead times) that are omitted from the main text for space. All tables in this appendix use the forecast protocol of Section 3 (input pair spaced by the evaluated horizon); the λ-sweeps of Appendices E.6 and F use the sweep protocol (input pair pinned to 3 h) and are not directly comparable in absolute value. The two samples agree on every ranking: on the subset, LWI has lower RMSE than every grid-space interpolant on all four near-surface variables at both horizons and the chain leads on Z500; the one difference is Z500 against the cubic spline, where the full-year evaluation puts the spline slightly ahead of LWI (0.9 and $0.8\,\mathrm{m}^2/\mathrm{s}^2$) whereas the subset had LWI marginally ahead.

C.1 Complete baseline set on the 64-initialization subset

Table 6: RMSE (lower is better) at lead times 3–6 h on Aurora, mean ± std over the random 64-initialization subset of ERA5 2022 (forecast protocol). $\mathcal{W}^{\mathrm{Aur}}_{\Delta t}$ is the natively fine-tuned anchor at lead time Δt; $\mathcal{W}^{\mathrm{Aur}}_{\lambda} = (1-\lambda)\mathcal{W}^{\mathrm{Aur}}_{3\mathrm{h}} + \lambda\mathcal{W}^{\mathrm{Aur}}_{6\mathrm{h}}$ ($\lambda = 1/3$ for 4 h, $\lambda = 2/3$ for 5 h). *Chain (Aur)* uses $\mathcal{W}^{\mathrm{Aur}}_{1\mathrm{h}} \to \mathcal{W}^{\mathrm{Aur}}_{3\mathrm{h}}$ at 4 h and $\mathcal{W}^{\mathrm{Aur}}_{1\mathrm{h}} \to \mathcal{W}^{\mathrm{Aur}}_{1\mathrm{h}} \to \mathcal{W}^{\mathrm{Aur}}_{3\mathrm{h}}$ at 5 h. $(\mathcal{W}^{\mathrm{Aur}}_{3\mathrm{h}})^{\times}$ and $(\mathcal{W}^{\mathrm{Aur}}_{6\mathrm{h}})^{\times}$ are the native models asked to output the wrong horizon. Grid-space baselines (persistence, linear tendency, linear/quadratic/cubic output interpolation) are computed from the 3h and 6h forecasts; the quadratic fit and the cubic spline coincide with only two anchor forecasts. **Bold**: best within each lead-time group. Underlined: second-best.

Lead (h)	Method / Setup	RMSE ↓			
		Z500	T2M	T850	U10
3	$\mathcal{W}^{\mathrm{Aur}}_{3\mathrm{h}}$	12.988 ± 0.616	0.441 ± 0.032	0.241 ± 0.007	0.276 ± 0.010
4	Persistence	173.237 ± 12.183	2.165 ± 0.311	1.227 ± 0.085	1.723 ± 0.077
	Linear tendency	38.498 ± 4.475	0.825 ± 0.128	0.473 ± 0.081	0.665 ± 0.100
	Linear output	26.768 ± 5.108	0.632 ± 0.069	0.372 ± 0.075	0.494 ± 0.091
	Quadratic fit	23.816 ± 6.107	0.654 ± 0.095	0.372 ± 0.081	0.492 ± 0.102
	Cubic spline	23.816 ± 6.107	0.654 ± 0.095	0.372 ± 0.081	0.492 ± 0.102
	LWI (ours) $\mathcal{W}^{\mathrm{Aur}}_{\lambda=1/3}$	23.709 ± 4.268	$\mathbf{0.613 \pm 0.069}$	$\mathbf{0.340 \pm 0.070}$	$\mathbf{0.407 \pm 0.097}$
	Chain (Aur)	$\mathbf{19.808 \pm 7.979}$	0.657 ± 0.072	0.342 ± 0.098	0.409 ± 0.133
	$(\mathcal{W}^{\mathrm{Aur}}_{3\mathrm{h}})^{\times}$	48.731 ± 4.171	0.784 ± 0.239	0.456 ± 0.077	0.603 ± 0.102
	$(\mathcal{W}^{\mathrm{Aur}}_{6\mathrm{h}})^{\times}$	101.943 ± 6.303	1.277 ± 0.227	0.819 ± 0.047	1.064 ± 0.054
5	Persistence	205.961 ± 15.237	2.672 ± 0.246	1.432 ± 0.096	1.981 ± 0.089
	Linear tendency	77.606 ± 3.929	1.210 ± 0.090	0.771 ± 0.061	1.163 ± 0.082
	Linear output	25.796 ± 3.291	0.540 ± 0.022	0.372 ± 0.059	0.492 ± 0.073
	Quadratic fit	24.022 ± 3.826	0.570 ± 0.027	0.383 ± 0.061	0.512 ± 0.077
	Cubic spline	24.022 ± 3.826	0.570 ± 0.027	0.383 ± 0.061	0.512 ± 0.077
	LWI (ours) $\mathcal{W}^{\mathrm{Aur}}_{\lambda=2/3}$	23.944 ± 3.570	$\mathbf{0.511 \pm 0.043}$	$\mathbf{0.352 \pm 0.056}$	$\mathbf{0.409 \pm 0.078}$
	Chain (Aur)	$\mathbf{19.453 \pm 5.184}$	0.531 ± 0.038	0.355 ± 0.061	0.415 ± 0.084
	$(\mathcal{W}^{\mathrm{Aur}}_{3\mathrm{h}})^{\times}$	97.031 ± 6.116	1.415 ± 0.109	0.725 ± 0.065	1.008 ± 0.063
	$(\mathcal{W}^{\mathrm{Aur}}_{6\mathrm{h}})^{\times}$	59.443 ± 3.587	0.673 ± 0.057	0.575 ± 0.042	0.675 ± 0.053
6	$\mathcal{W}^{\mathrm{Aur}}_{6\mathrm{h}}$	18.706 ± 3.828	0.451 ± 0.034	0.353 ± 0.061	0.417 ± 0.082

Table 7: ACC at lead times 3–6 h on the 64-initialization subset of ERA5 2022. Method notation as in Table 6. ACC is essentially saturated for Z500 across viable methods, so RMSE is the discriminating point-wise metric in the main text; LWI nevertheless leads on T2m, T850 and U10.

Lead (h)	Method / Setup	ACC ↑			
		Z500	T2M	T850	U10
3	$\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}}$	0.999 ± 0.000	0.989 ± 0.002	0.997 ± 0.000	0.997 ± 0.000
4	Persistence	0.976 ± 0.003	0.805 ± 0.064	0.939 ± 0.009	0.901 ± 0.012
	Linear tendency	0.998 ± 0.000	0.956 ± 0.015	0.990 ± 0.003	0.985 ± 0.004
	Linear output	0.999 ± 0.000	$\underline{0.973 \pm 0.008}$	$\underline{0.994 \pm 0.002}$	0.991 ± 0.003
	Quadratic fit	0.999 ± 0.000	0.971 ± 0.010	0.993 ± 0.002	0.991 ± 0.003
	Cubic spline	0.999 ± 0.000	0.971 ± 0.010	0.993 ± 0.002	0.991 ± 0.003
	LWI (ours)	0.999 ± 0.000	$\mathbf{0.974 \pm 0.007}$	$\mathbf{0.995 \pm 0.002}$	$\mathbf{0.994 \pm 0.003}$
	Chain (Aur)	0.999 ± 0.000	0.971 ± 0.009	0.994 ± 0.003	$\underline{0.993 \pm 0.004}$
	$(\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}})^{\times}$	0.998 ± 0.000	0.958 ± 0.026	0.991 ± 0.003	0.987 ± 0.004
	$(\mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}})^{\times}$	0.991 ± 0.001	0.882 ± 0.042	0.971 ± 0.005	0.961 ± 0.005
5	Persistence	0.967 ± 0.005	0.684 ± 0.065	0.916 ± 0.012	0.869 ± 0.016
	Linear tendency	0.995 ± 0.000	0.908 ± 0.022	0.976 ± 0.004	0.959 ± 0.007
	Linear output	0.999 ± 0.000	0.977 ± 0.004	0.994 ± 0.002	0.991 ± 0.002
	Quadratic fit	0.999 ± 0.000	0.975 ± 0.004	0.993 ± 0.002	0.990 ± 0.003
	Cubic spline	0.999 ± 0.000	0.975 ± 0.004	0.993 ± 0.002	0.990 ± 0.003
	LWI (ours)	0.999 ± 0.000	$\mathbf{0.980 \pm 0.004}$	$\mathbf{0.994 \pm 0.001}$	$\mathbf{0.994 \pm 0.002}$
	Chain (Aur)	0.999 ± 0.000	$\underline{0.978 \pm 0.004}$	$\underline{0.994 \pm 0.002}$	$\underline{0.993 \pm 0.002}$
	$(\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}})^{\times}$	0.992 ± 0.001	0.879 ± 0.026	0.978 ± 0.004	0.966 ± 0.005
	$(\mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}})^{\times}$	0.997 ± 0.000	0.965 ± 0.007	0.986 ± 0.002	0.984 ± 0.003
6	$\mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}}$	0.999 ± 0.000	0.984 ± 0.002	0.994 ± 0.002	0.993 ± 0.002

C.2 SEASONAL STRATIFICATION OF THE FULL-YEAR EVALUATION

Table 8: Seasonal stratification of the full-year 2022 evaluation (every 6-hourly initialization; DJF, MAM, JJA and SON contain about 365 initializations each). RMSE, mean ± std, of LWI and of the autoregressive chain on Z500 (m^2/s^2) and T2m (K) at 4 h and 5 h. The ranking of Table 2 holds in every season: the chain leads on Z500, LWI leads or ties on T2m.

	Z500, 4 h		Z500, 5 h		T2m, 4 h		T2m, 5 h	
Season	LWI	Chain	LWI	Chain	LWI	Chain	LWI	Chain
DJF	25.57 ± 4.34	22.41 ± 8.99	26.41 ± 4.36	21.98 ± 6.72	0.57 ± 0.06	0.62 ± 0.08	0.53 ± 0.03	0.53 ± 0.02
MAM	25.75 ± 4.10	20.87 ± 8.65	26.09 ± 4.34	21.43 ± 6.13	0.62 ± 0.07	0.63 ± 0.09	0.53 ± 0.04	0.53 ± 0.04
JJA	26.09 ± 4.81	23.85 ± 7.79	27.40 ± 5.04	22.98 ± 5.87	0.59 ± 0.05	0.67 ± 0.07	0.52 ± 0.04	0.53 ± 0.03
SON	26.46 ± 4.33	17.51 ± 6.86	26.31 ± 4.46	18.10 ± 4.65	0.59 ± 0.08	0.62 ± 0.07	0.51 ± 0.03	0.54 ± 0.04

D STATISTICAL SIGNIFICANCE: PAIRED BOOTSTRAP

We quantify the uncertainty of method-vs-method RMSE comparisons with a paired bootstrap over initialization times. For every (variable, lead-time) cell we compute the per-initialization latitude-weighted RMSE of each method, yielding length-N vectors $\{r_k^{(A)}\}_{k=1}^N$ and $\{r_k^{(B)}\}_{k=1}^N$ aligned by initialization k. We then draw $B = 10{,}000$ bootstrap samples by resampling the index set $\{1, \dots, N\}$ with replacement, compute $\bar{r}_b^{(A)} - \bar{r}_b^{(B)}$ on each resample, and report the 2.5 and 97.5 percentiles as a 95% confidence interval on $\Delta\mathrm{RMSE}_{A-B}$. A is declared significantly better than B when the upper end of this interval is below zero. Table 9 gives the intervals on the full-year protocol ($N = 1460$) for the two comparisons that matter in the main text; every significance statement in the text refers to this table.

Table 9: Paired-bootstrap 95% intervals on $\hat{\Delta} = \bar{r}_{\mathrm{LWI}} - \bar{r}_{\mathrm{B}}$ over every 6-hourly initialization of 2022 ($N = 1460$, $B = 10{,}000$, percentile intervals, written as center ± half-width). Negative $\hat{\Delta}$ means LWI is better. Every interval excludes zero except the one marked n.s.

Lead	Method B	Z500	T2m	T850	U10	V10
4 h	Chain 1h + 3h	$+4.81 \pm 0.19$	-0.04 ± 0.006	-0.01 ± 0.002	-0.02 ± 0.002	-0.02 ± 0.002
4 h	Cubic spline	$+0.86 \pm 0.11$	-0.02 ± 0.008	-0.02 ± 0.001	-0.06 ± 0.001	-0.10 ± 0.001
5 h	Chain 1h + 1h + 3h	$+5.43 \pm 0.10$	-0.01 ± 0.001	-0.01 ± 0.001	-0.02 ± 0.001	0.00 (n.s.)
5 h	Cubic spline	$+0.83 \pm 0.06$	-0.08 ± 0.003	-0.02 ± 0.001	-0.08 ± 0.001	-0.10 ± 0.001

E MECHANISTIC ANALYSIS

This section reports the diagnostics that are summarized in Section 3. The unifying narrative is the asymmetry between the two adaptation pairs: $(\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}}, \mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}})$ is a small, low-rank, high-CKA, comparable-receptive-field pair, whereas the 1 h adaptation has restructured into a different low-loss basin despite being trained from the same initialization under an identical recipe.

E.1 WEIGHT-DELTA GEOMETRY

Table 10 reports per-component relative L2 norm, cosine similarity, and L_∞ norm of the weight deltas $\Delta\mathcal{W}_{a,b} = \mathcal{W}_a - \mathcal{W}_b$. The qualitative pattern is consistent across all backbone stages and the encoder/decoder: the 3h vs. 6h relative L2 is a factor of 1.4–3.4 smaller than 1h vs. 6h on every block (e.g. BB↑3: 0.0051 vs. 0.0172; encoder: 0.0387 vs. 0.0521). Cosine similarities are essentially 1 everywhere, so the difference is in magnitude rather than direction. The 1h vs. 3h column is closer to 1h vs. 6h than to 3h vs. 6h on every block except the encoder, again indicating that the 1h adaptation is the outlier of the three. In raw parameter geometry, the 3h and 6h checkpoints sit in the same neighborhood and the 1h checkpoint sits substantially apart.

Table 10: Mean relative L2 norm, mean cosine similarity, and maximum L_∞ norm of weight deltas between Aurora checkpoints fine-tuned at different lead times, broken down by architectural component. Cosine similarities are raw aggregates over per-parameter pairs.

	1 h vs. 6 h			3 h vs. 6 h			1 h vs. 3 h		
Component	Rel. L2	Cos. sim.	L_∞	Rel. L2	Cos. sim.	L_∞	Rel. L2	Cos. sim.	L_∞
Encoder	0.0521	0.9960	0.0493	0.0387	0.9973	0.0080	0.0373	0.9975	0.0474
BB↓1	0.0164	0.9998	0.0344	0.0080	1.0000	0.0090	0.0150	0.9999	0.0271
BB↓2	0.0221	0.9997	0.0364	0.0126	0.9999	0.0155	0.0199	0.9998	0.0409
BB↓3	0.0228	1.0000	0.0239	0.0159	1.0000	0.0101	0.0233	1.0000	0.0216
BB↑1	0.0206	1.0000	0.0332	0.0148	1.0000	0.0127	0.0200	1.0000	0.0317
BB↑2	0.0222	0.9997	0.0370	0.0101	1.0000	0.0117	0.0192	0.9998	0.0349
BB↑3	0.0172	0.9998	0.0373	0.0051	1.0000	0.0088	0.0156	0.9998	0.0338
Decoder	0.0262	0.9992	0.0423	0.0112	0.9999	0.0075	0.0207	0.9995	0.0384
Other	0.0373	0.9992	0.0039	0.0230	0.9997	0.0025	0.0387	0.9990	0.0045

Table 11: Spectral decomposition statistics of weight deltas $\Delta W = W_a - W_b$ between Aurora checkpoints fine-tuned at different lead times, aggregated by architectural stage. "Eff. rank" is the Roy effective rank $(\sum \sigma)^2 / \sum \sigma^2$; "rank@95%" is the number of singular values needed to capture 95% of the Frobenius energy; "top-1 frac." is $\sigma_1 / \sum \sigma$.

	1 h vs. 6 h			3 h vs. 6 h		
Stage	Eff. rank	rank@95%	Top-1 frac.	Eff. rank	rank@95%	Top-1 frac.
Encoder	85.1	44.0	0.169	82.3	46.0	0.169
BB↓1	131.0	132.0	0.154	117.3	153.0	0.191
BB↓2	252.4	266.0	0.156	265.9	286.0	0.184
BB↓3	345.6	303.0	0.136	464.4	425.5	0.166
BB↑1	369.3	340.0	0.131	444.4	415.0	0.168
BB↑2	264.2	252.5	0.153	279.3	282.5	0.167
BB↑3	148.4	129.5	0.146	141.7	125.0	0.153
Decoder	141.1	77.0	0.044	149.2	91.0	0.041

E.2 Spectral structure of the weight deltas

Tables 11 and 12 report the SVD spectrum of each $\Delta \mathcal{W}$ aggregated by stage. The Roy effective rank by itself is similar across pairs and large across all blocks, which on its own does *not* discriminate the 1h–6h delta from the 3h–6h delta. The cleanest discriminator is the median rank required to capture 50% of the Frobenius energy (Table 12): for the 3h–6h pair, only ∼3–28 singular values per stage suffice (e.g. Encoder 3, BB↓1 7, BB↑3 13.5), whereas for 1h–6h the corresponding numbers are ∼4–32 but with a markedly more diffuse top-1 fraction across stages. The 3h–6h delta is, layer for layer, dominated by a handful of directions, which is the regime in which a single scalar interpolation in weight space is well posed and in which the task-vector view of Section 2 applies.

E.3 Fisher- vs. raw-L2 layer rankings

Table 13 reports the Spearman rank correlation between layer rankings by raw L2 distance and by Fisher-weighted distance for each Aurora pair, together with the size of the top-5 overlap. Two observations are worth flagging. First, the rank correlations are uniformly high ($\rho \approx 0.65$–0.73, $p < 0.001$ in every cell), so the two rankings broadly agree at the layer level. Second, the top-5 overlap is exactly zero in every pair: Fisher reweighting promotes `reduction.weight`, `surf_head.weight` and `to_out.weight` into the top tier, while raw L2 places `fc1.weight` layers at the top. This is consistent with Fisher reweighting up-weighting the loss-sensitive output and dimensionality-reducing modules. For our purposes, this corroborates the main-text statement that Fisher-weighted interpolation does not change the conclusion: the high rank correlation makes the two views agree on which *stages* are central, all of these layers are interpolated anyway, and Section E.8 shows that freezing any subset degrades the result, so the reweighting has nothing substantive to change.

Table 12: Spectral decomposition statistics of weight deltas $\Delta W = W_a - W_b$ between Aurora checkpoints fine-tuned at different lead times, aggregated by architectural stage. "Eff. rank" is the Roy effective rank $(\sum \sigma)^2 / \sum \sigma^2$; "rank@50%" is the number of singular values needed to capture 50% of the Frobenius energy; "top-1 frac." is $\sigma_1 / \sum \sigma$.

	1 h vs. 6 h			3 h vs. 6 h		
Stage	Eff. rank	rank@50%	Top-1 frac.	Eff. rank	rank@50%	Top-1 frac.
Encoder	85.1	4.0	0.169	82.3	3.0	0.169
BB↓1	131.0	12.0	0.154	117.3	7.0	0.191
BB↓2	252.4	22.0	0.156	265.9	20.0	0.184
BB↓3	345.6	11.5	0.136	464.4	13.0	0.166
BB↑1	369.3	14.5	0.131	444.4	26.0	0.168
BB↑2	264.2	32.0	0.153	279.3	28.5	0.167
BB↑3	148.4	18.0	0.146	141.7	13.5	0.153
Decoder	141.1	9.5	0.044	149.2	9.0	0.041

Table 13: Spearman rank correlation between layer rankings by raw L2 distance and Fisher-weighted distance for each Aurora model pair. "Overlap" counts how many of the top-5 layers by raw L2 also appear in the top-5 by Fisher distance. The top-5 layers by raw L2 are dominated by `fc1.weight` across all pairs; the top-5 by Fisher distance include `reduction.weight`, `fc1.weight`, `surf_head.weight`, and `to_out.weight` (overlap = 0 in all cases).

Pair	ρ	p-value	Overlap
1 h vs. 6 h	0.669	< 0.001	0 / 5
3 h vs. 6 h	0.652	< 0.001	0 / 5
1 h vs. 3 h	0.726	< 0.001	0 / 5

E.4 CKA PER STAGE

Table 14 reports mean diagonal CKA per architectural stage for all Aurora pairs. Three regimes are visible. Encoder, BB↓1, BB↓2 and BB↑3 are essentially saturated (≥ 0.97) for the 3h–6h pair, indicating near-identical representations; the 1h–6h and 1h–3h columns drop to ≤ 0.06 at the same stages, a striking representational gap that nevertheless coexists with ≥ 0.99 cosine similarity in raw weights, a reminder that representation similarity and weight similarity are different objects. Mid-depth blocks (BB↓3, BB↑1, BB↑2) show a partial CKA drop *even* for the 3h–6h pair (down to 0.55–0.74), which we read as the structural signature of the receptive-field shift discussed in the main text: the longest-horizon attention reach lives in the mid blocks. The same mid-depth dip appears on Pangu-Weather and GraphCast (Appendix F).

The CKA picture is consistent with orthogonal-Procrustes distances on the same layers: Table 15 reports the Spearman correlation between CKA and Procrustes layer rankings, which is essentially 1 for the 3h–6h pair ($\rho = 0.996$) and notably lower (still 0.87) for the 1h–6h pair, again with the 1h adaptation as the outlier. Together with the L2 evidence in Section E.1, these diagnostics jointly identify the 3h–6h pair as the natural interpolable segment and the 1h adaptation as a separate basin.

E.5 EFFECTIVE RECEPTIVE FIELD

Table 16 reports the effective receptive field (ERF) radius, defined as the median great-circle distance to grid cells with $|\partial Z_{500}/\partial x| > 0.1 \max$, for the three Aurora variants. The progression is monotone in lead time: $\mathcal{W}_{1\mathrm{h}}^{\mathrm{Aur}} \to 350$ km, $\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}} \to 390$ km, $\mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}} \to 557$ km. Two takeaways. First, the gap from 3 h to 6 h ($\sim$170 km) is concentrated in the synoptic regime, consistent with the dominant length scale a 6 h forecast has to integrate over; the gap from 1 h to 3 h ($\sim$40 km) is much smaller. Second, the ratio $557/350 \approx 1.6$ is well below $\sqrt{6/1} \approx 2.45$, which we tentatively associate with the fact that ERA5-derived dynamics are not isotropic random walks. The qualitative ordering, and the asymmetry

Table 14: Mean diagonal CKA similarity per architectural stage for all model pairs. Higher values indicate more similar representations between the two lead-time variants at that stage.

		Model pair		
Architecture	Stage	1 h vs. 6 h	3 h vs. 6 h	1 h vs. 3 h
Aurora	Encoder	0.056	1.000	0.055
	BB-down-1	0.039	1.000	0.038
	BB-down-2	0.064	0.974	0.063
	BB-down-3	0.123	0.562	0.137
	BB-up-1	0.157	0.545	0.181
	BB-up-2	0.111	0.739	0.113
	BB-up-3	0.043	0.998	0.042

Table 15: Spearman rank correlation between diagonal CKA similarity and Procrustes distance rankings across all model pairs. $\rho \approx 1$ indicates the two metrics agree on layer ordering; values near 0 indicate disagreement warranting further investigation.

Pair	1 h vs 3 h	1 h vs 6 h	3 h vs 6 h
Aurora	0.974 (p<0.001)	0.873 (p<0.001)	0.996 (p<0.001)

in step sizes, is what one would predict if the network is solving a Δt-dependent advection problem in which the dominant correlation length grows sub-linearly with Δt.

E.6 Low-loss valley along the 3h–6h interpolation path (fixed-target sweep)

Setup. For Aurora, define $W_\lambda = (1-\lambda)\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}} + \lambda \mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}}$ and $\Delta W = \mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}} - \mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}}$. We evaluate $L(W_\lambda)$ at a *fixed target horizon* $T^\star$ for every λ on a grid of step 0.05 (Z500 RMSE on the 64-initialization subset of 2022). *Protocol.* For every λ, endpoints included, the input pair is spaced by the 3h anchor's stencil, $(t-3\mathrm{h}, t)$, while the lead-time conditioning and the target offset are set to $T^\star$. This differs from the forecast protocol of the tables, where the input pair is spaced by the evaluated horizon itself, so absolute values are not comparable between the two: $\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}}$ at $T^\star = 4\,\mathrm{h}$ scores 44.94 here and 48.7 under the forecast protocol (Table 6), and LWI at the nominal λ scores lower under the forecast protocol than the sweep minimum here. The sweeps characterize the loss surface along the path; the tables characterize the forecasts. The smoothness inequality

$$L(W_\lambda) \le (1-\lambda) L(\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}}) + \lambda L(\mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}}) + \frac{\kappa}{2}\lambda(1-\lambda)\|\Delta W\|^2 \tag{6}$$

is included only as motivating context: with the fixed-$T^\star$ protocol the realized loss undercuts the linear baseline at every interior λ, so the bound is satisfied vacuously with $\kappa = 0$. The substantive finding is the existence and location of the interior minimum $\lambda^\star(T^\star)$ and the size of the gap to the convex-combination baseline.

Results at $T^\star = 4\,\mathrm{h}$. See Table 17 and Figure 3.

Results at $T^\star = 5\,\mathrm{h}$. See Table 18 and Figure 4.

Reading. (i) The empirical loss is *below* the linear baseline at every interior λ for the protocol above. (ii) The largest curvature consistent with the data is $\kappa_{\max} = 0$, so Equation (6) is vacuously satisfied and is not the substantive content of this section. (iii) The substantive content is the existence of an interior minimum and its empirical location $\lambda^\star(T^\star)$, which we interpret as a low-loss valley in the lead-time dimension of the Aurora weight space. The endpoint separation $\|\Delta W\|_F \approx 45.7$ is small relative to typical Aurora weight norms.

Table 16: Effective Receptive Field radius (km) for each Aurora variant, computed as the median great-circle distance to grid cells where $|\partial Z_{500}/\partial x| > 0.1 \cdot \max$ over $N = 50$ test samples. Per-sample standard deviations are omitted because the underlying per-init. ERF estimator is noisy at this sample size; we report only the median across the $N = 50$ initializations.

Model	ERF radius (km)	N
1h	350.3	50
3h	389.6	50
6h	557.1	50

Table 17: Smoothness-bound diagnostics for the 3h→6h Aurora interpolation, evaluated at fixed target horizon $T^\star = 4$ h on Z500. $L_{\text{lin}}(\lambda) = (1-\lambda)L(\mathcal{W}_{3\text{h}}^{\text{Aur}}) + \lambda L(\mathcal{W}_{6\text{h}}^{\text{Aur}})$ is the convex-combination baseline; $\varepsilon(\lambda) = L(W_\lambda) - L_{\text{lin}}(\lambda)$ is the excess error along the path. Both endpoint models and all interpolated models are evaluated on the same task (same target horizon, same time-conditioning value $T^\star$), as required for the bound to be a valid statement about a single loss surface.

λ	$T^\star$ (h)	$L(W_\lambda)$	$L_{\text{lin}}(\lambda)$	$\varepsilon(\lambda)$
0.000	4	44.94	44.94	0.00
0.050	4	40.15	47.83	−7.67
0.100	4	35.58	50.72	−15.14
0.150	4	31.30	53.61	−22.31
0.200	4	27.63	56.50	−28.87
0.250	4	25.22	59.39	−34.17
0.300	4	**24.75**	**62.28**	**−37.53**
0.350	4	26.42	65.17	−38.75
0.400	4	29.75	68.06	−38.31
0.450	4	34.19	70.95	−36.76
0.500	4	39.36	73.84	−34.48
0.550	4	45.03	76.73	−31.70
0.600	4	50.99	79.62	−28.63
0.650	4	57.12	82.51	−25.39
0.700	4	63.39	85.40	−22.01
0.750	4	69.75	88.29	−18.54
0.800	4	76.20	91.18	−14.98
0.850	4	82.72	94.07	−11.35
0.900	4	89.31	96.96	−7.66
0.950	4	95.97	99.85	−3.88
1.000	4	102.74	102.74	0.00

$L(\mathcal{W}_{3\text{h}}^{\text{Aur}}) = 44.94$, $L(\mathcal{W}_{6\text{h}}^{\text{Aur}}) = 102.74$, $\|\Delta W\|_F = 45.7$, $\kappa_{\max} = 0$.

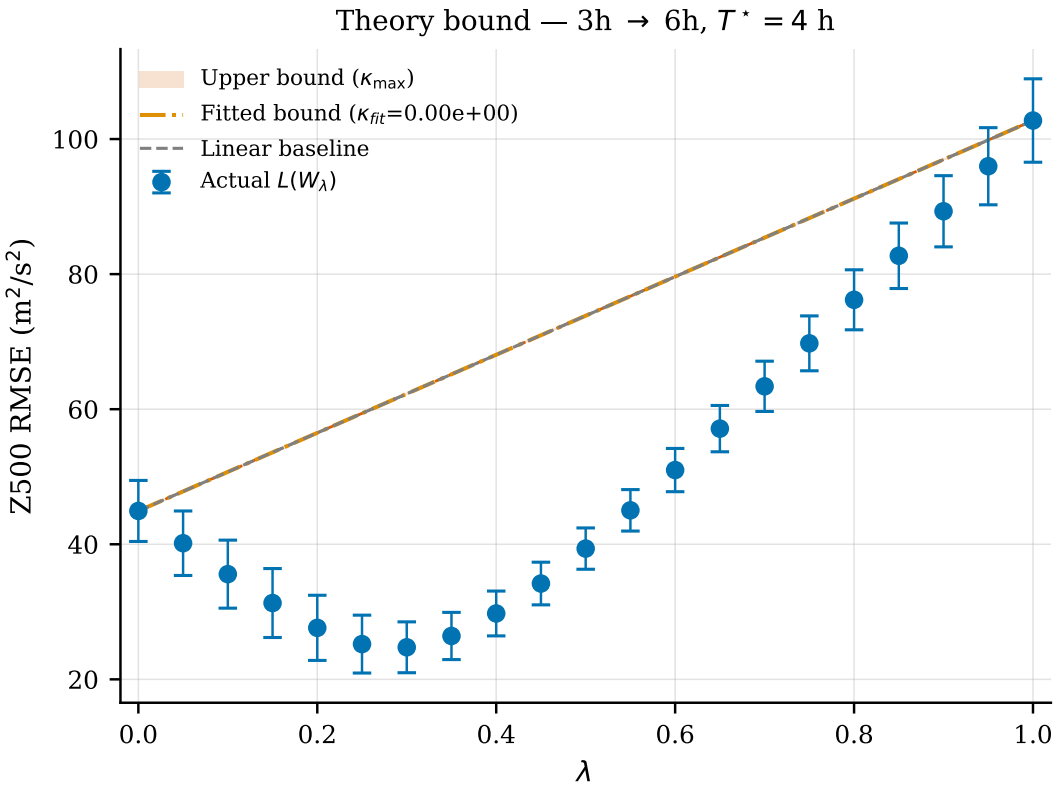


Figure 3: **Smoothness-bound diagnostics at fixed target horizon $T^\star = 4$ h (Z500).** Z500 RMSE of the linearly interpolated checkpoint $W_\lambda = (1-\lambda)W_{3\mathrm{h}} + \lambda W_{6\mathrm{h}}$ (blue dots, $\pm 1\sigma$ bootstrap over the 64-initialization subset of 2022, sweep protocol) versus the convex-combination baseline $L_{\mathrm{lin}}(\lambda)$ (dashed grey) and the fitted Lipschitz-style upper bound (orange dash-dot, $\kappa_{\mathrm{fit}} = 0$). The empirical curve lies strictly below the chord at every interior λ, with minimum $L_{\mathrm{min}} \approx 24.75\,\mathrm{m}^2/\mathrm{s}^2$ at $\lambda^\star \approx 0.30$, 45% below the better endpoint and 60% below the chord at $\lambda^\star$. The vanishing fitted curvature ($\kappa_{\mathrm{fit}} = 0$) confirms that the smoothness inequality is not active; the substantive finding is the low-loss valley itself.

Table 18: Smoothness-bound diagnostics for the 3h→6h Aurora interpolation, evaluated at fixed target horizon $T^\star = 5\,\mathrm{h}$ on Z500. Notation as in Table 17.

λ	$T^\star$ (h)	$L(W_\lambda)$	$L_{\mathrm{lin}}(\lambda)$	$\varepsilon(\lambda)$
0.000	5	87.66	87.66	0.00
0.050	5	82.13	86.80	−4.67
0.100	5	76.49	85.94	−9.45
0.150	5	70.75	85.09	−14.34
0.200	5	64.83	84.23	−19.40
0.250	5	58.58	83.37	−24.79
0.300	5	52.36	82.51	−30.15
0.350	5	46.82	81.66	−34.83
0.400	5	42.04	80.80	−38.76
0.450	5	38.03	79.94	−41.91
0.500	5	34.98	79.09	−44.11
0.550	5	33.20	78.23	−45.03
0.600	5	**32.92**	**77.37**	**-44.46**
0.650	5	34.15	76.51	−42.36
0.700	5	36.74	75.66	−38.92
0.750	5	40.41	74.80	−34.38
0.800	5	44.91	73.94	−29.03
0.850	5	50.01	73.09	−23.07
0.900	5	55.57	72.23	−16.66
0.950	5	64.49	71.37	−6.88
1.000	5	70.51	70.51	0.00

$L(\mathcal{W}_{3\mathrm{h}}^{\mathrm{Aur}}) = 87.66$, $L(\mathcal{W}_{6\mathrm{h}}^{\mathrm{Aur}}) = 70.51$, $\|\Delta W\|_F = 45.7$, $\kappa_{\max} = 0.$

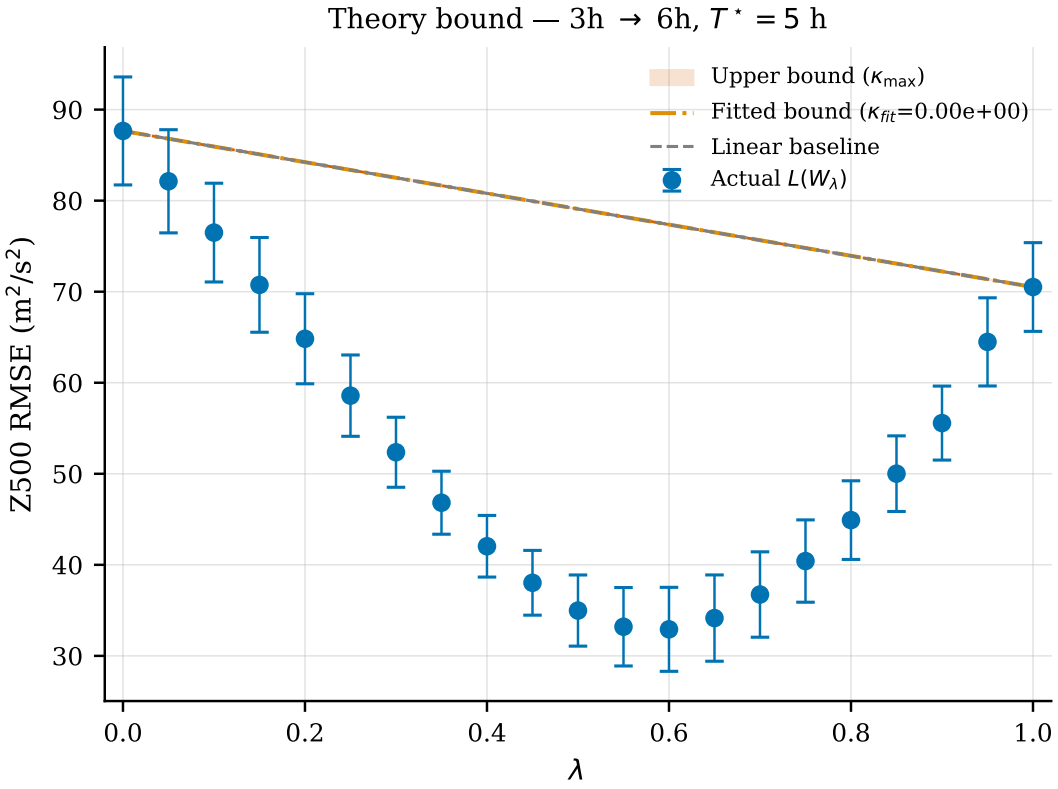


Figure 4: **Smoothness-bound diagnostics at fixed target horizon $T^\star = 5$ h (Z500).** Same protocol as Figure 3 (64-initialization subset of 2022, full λ-sweep, error bars are bootstrap 1σ), but every interpolated checkpoint is evaluated at $T^\star = 5$ h. The interpolation curve again lies strictly below the chord, with optimum $L_{\min} \approx 32.92\,\mathrm{m}^2/\mathrm{s}^2$ at $\lambda^\star \approx 0.60$, 53% below the better endpoint and 57% below the chord at $\lambda^\star$. The fitted curvature vanishes ($\kappa_{\mathrm{fit}} = 0$), confirming that the smoothness inequality is not active anywhere on the path; the qualitative picture is unchanged from $T^\star = 4$ h.

E.7 SEED CONTROL

Table 19: Seed control for the 3h–6h connectivity on Aurora. The 3h adaptation was fine-tuned with three independent seeds under the identical recipe; for each, the stage-wise CKA against the 6h checkpoint, the fixed-$T^\star$ sweep at $T^\star = 4\,\mathrm{h}$ (Z500, 64 initializations, protocol of Appendix E.6) and the valley depth relative to the better wrong-horizon endpoint were recomputed. Seed 0 is the checkpoint used everywhere else in the paper.

Seed	CKA(3h, 6h) ≥ 0.97	$\lambda^\star(4\mathrm{h})$	$L(W_{\lambda^\star})$	depth vs. better endpoint
0 (reference)	4 of 7 (Table 14)	0.30	24.75	45%
1	4 of 7	0.30	25.1	44%
2	5 of 7	0.30	24.5	45%
mean $\pm$ std	4–5 of 7	0.30 ± 0.00	24.8 ± 0.3	$45 \pm 1\%$

E.8 INTERPOLATION OF WEIGHT SUBSETS

We additionally compared interpolation of all weights against interpolation of only the (Fisher-identified) inert subsets, while keeping the loss-sensitive layers fixed at one endpoint. The all-weights linear scheme wins uniformly; partial-subset interpolation degrades with the size of the frozen complement, with the most pronounced degradation when the loss-sensitive output heads are not interpolated (Figure 5).

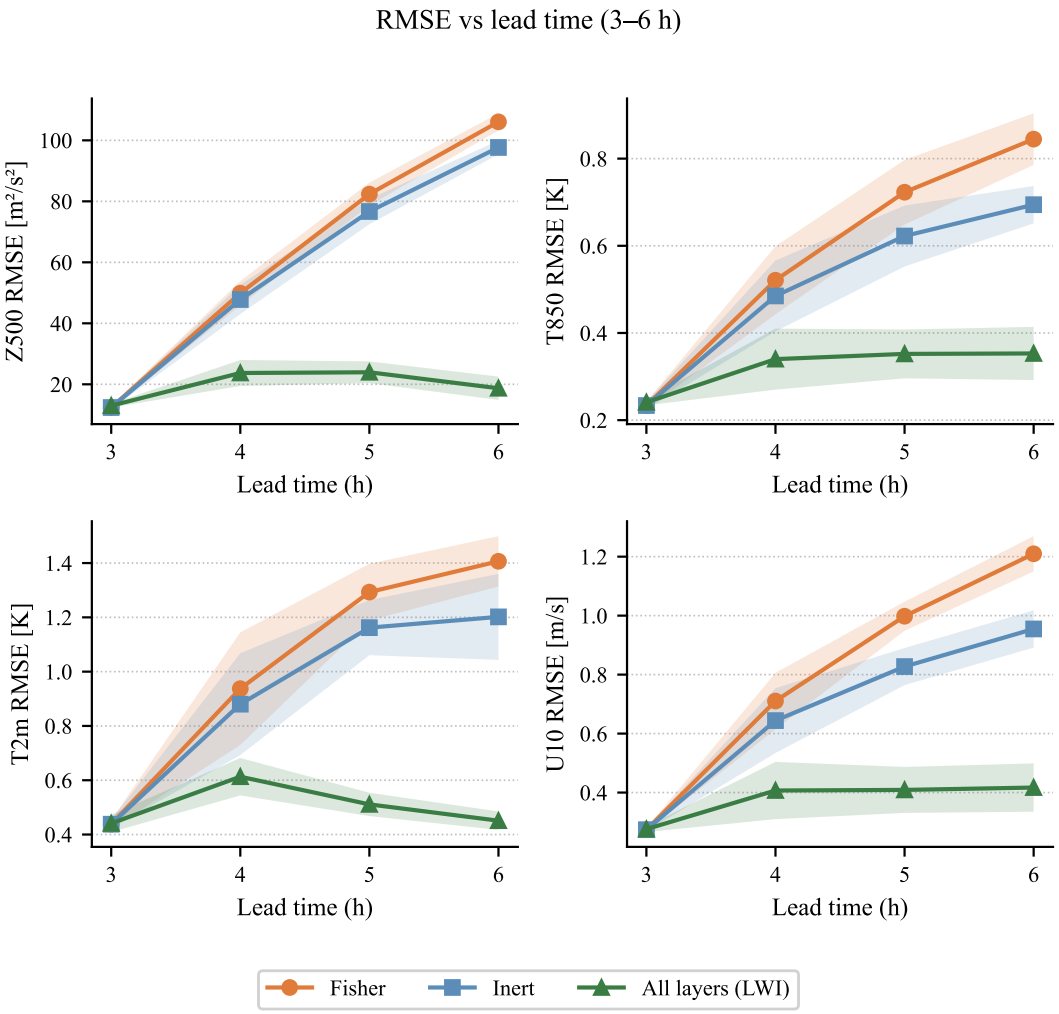


Figure 5: **RMSE vs. lead time for weight-subset interpolation ablations.** Mean RMSE $\pm$ 1 std over $N = 4$ ERA5 test samples at lead times 3–6 h on Aurora. *All layers (LWI)* interpolates all weights linearly ($\mathcal{W}_\lambda = (1-\lambda)\mathcal{W}_{3\text{h}}^{\text{Aur}} + \lambda\mathcal{W}_{6\text{h}}^{\text{Aur}}$); *Fisher* interpolates only the Fisher-identified loss-sensitive layers while freezing the remainder at the 3 h endpoint; *Inert* interpolates only the complementary low-Fisher layers while freezing the loss-sensitive heads at the 3 h endpoint. All-weights interpolation wins uniformly; partial-subset schemes degrade progressively, with the most pronounced degradation when the output heads are not interpolated (*Inert*).

F Pangu-Weather and GraphCast

This appendix tabulates the cross-architecture experiments of Section 4.1. All sweeps use the fixed-$T^\star$ protocol of Appendix E.6 (Z500 RMSE in m^2/s^2; L_{lin} is the chord between the endpoint losses and $\varepsilon(\lambda) = L(W_\lambda) - L_{\text{lin}}(\lambda)$), evaluated on 64 random initializations of 2022. As for Aurora, the input stencil of the two-state model (GraphCast) is pinned to the 3h anchor's spacing for every λ; Pangu-Weather takes a single input state. Because neither Pangu-Weather nor GraphCast has a lead-time input, "fixed $T^\star$" only fixes the target offset of the evaluation data; every model on the path is scored on literally the same task. The two Pangu-Weather sweeps were run as separate campaigns (the official pair through its released ONNX graphs, the fine-tuned pair through our training stack), so the official 6h endpoint takes different values in Tables 20 and 21 (98.18 and 86.21); comparisons are made within each sweep only.

F.1 PANGU-WEATHER: NEGATIVE CONTROL (INDEPENDENTLY TRAINED OFFICIAL CHECKPOINTS)

Table 20: Fixed-target sweep between the officially released, independently trained Pangu-Weather 3h and 6h models at $T^\star = 4\,\text{h}$. $\varepsilon(\lambda) > 0$ at every interior point: the interpolated models cross a loss barrier, peaking at $\lambda = 0.5$ at $27\times$ the better endpoint. At $\lambda = 1/3$ the T2m RMSE is about 7 K against 0.50 and 0.57 K for the endpoints. Endpoint separation $\|\Delta W\|_F = 4766.5$, against 45.7 for the Aurora pair and 10.7 for the fine-tuned Pangu-Weather pair of Table 21.

λ	$L(W_\lambda)$	$L_{\text{lin}}(\lambda)$	$\varepsilon(\lambda)$
0.000	57.30	57.30	0.00
0.050	293.01	59.34	+233.67
0.100	536.92	61.38	+475.54
0.150	778.02	63.43	+714.59
0.200	1012.23	65.47	+946.76
0.250	1193.61	67.52	+1126.09
0.300	1315.39	69.56	+1245.82
0.350	1417.56	71.61	+1345.95
0.400	1507.41	73.65	+1433.76
0.450	1561.97	75.70	+1486.27
0.500	**1562.51**	77.74	+1484.77
0.550	1518.17	79.78	+1438.38
0.600	1437.63	81.83	+1355.81
0.650	1332.81	83.87	+1248.94
0.700	1215.68	85.92	+1129.76
0.750	1075.30	87.96	+987.34
0.800	900.47	90.01	+810.46
0.850	695.35	92.05	+603.29
0.900	482.60	94.10	+388.50
0.950	263.76	96.14	+167.62
1.000	98.18	98.18	0.00

$L(W_{3\text{h}}) = 57.30$, $L(W_{6\text{h}}) = 98.18$, $\|\Delta W\|_F = 4766.5$.

F.2 PANGU-WEATHER: POSITIVE CONTROL (3H ANCHOR FINE-TUNED FROM THE OFFICIAL 6H WEIGHTS)

Table 21: Fixed-target sweeps between the fine-tuned Pangu-Weather 3h anchor and the official 6h model at $T^\star = 4\,\mathrm{h}$ and $T^\star = 5\,\mathrm{h}$. $\varepsilon(\lambda) < 0$ at every interior point of both sweeps; the interior minima are 30.58 at $\lambda^\star = 0.50$ (48% below the better endpoint, flat over $[0.45, 0.55]$) and 28.93 at $\lambda^\star = 0.80$ (38% below, flat over $[0.75, 0.80]$). $\|\Delta W\|_F = 10.7$, $\kappa_{\max} = 0$ for both.

	$T^\star = 4\,\mathrm{h}$			$T^\star = 5\,\mathrm{h}$		
λ	$L(W_\lambda)$	$L_{\text{lin}}(\lambda)$	$\varepsilon(\lambda)$	$L(W_\lambda)$	$L_{\text{lin}}(\lambda)$	$\varepsilon(\lambda)$
0.000	59.21	59.21	0.00	99.01	99.01	0.00
0.050	55.36	60.56	−5.20	94.80	96.40	−1.60
0.100	51.50	61.91	−10.41	90.46	93.79	−3.34
0.150	47.69	63.26	−15.57	85.98	91.18	−5.20
0.200	43.93	64.61	−20.68	81.36	88.57	−7.21
0.250	40.32	65.96	−25.64	76.58	85.96	−9.37
0.300	36.97	67.31	−30.34	71.70	83.35	−11.64
0.350	34.14	68.66	−34.52	66.71	80.74	−14.03
0.400	31.89	70.01	−38.12	61.60	78.13	−16.53
0.450	30.63	71.36	−40.73	56.40	75.51	−19.11
0.500	**30.58**	72.71	−42.13	51.16	72.90	−21.74
0.550	31.92	74.06	−42.14	45.98	70.29	−24.32
0.600	34.51	75.41	−40.90	40.97	67.68	−26.71
0.650	38.31	76.76	−38.45	36.34	65.07	−28.73
0.700	43.14	78.11	−34.98	32.40	62.46	−30.06
0.750	48.74	79.46	−30.73	29.72	59.85	−30.13
0.800	55.10	80.81	−25.71	**28.93**	57.24	−28.31
0.850	62.06	82.16	−20.11	30.49	54.63	−24.14
0.900	69.62	83.51	−13.89	34.25	52.01	−17.77
0.950	77.67	84.86	−7.20	39.83	49.40	−9.58
1.000	86.21	86.21	0.00	46.79	46.79	0.00

Table 22: Mean diagonal CKA per Pangu-Weather stage between the fine-tuned 3h anchor ("3h"), the official 6h model ("6h") and the nominal blends at $\lambda = 1/3$ ("4h") and $2/3$ ("5h"). The informative column is 3h vs. 6h, the analogue of the Aurora measurement in Table 14: seven of eight stages are at ≥ 0.976 and the minimum sits at mid-depth (ESL-3, 0.929). Columns pairing a blend with one of its own endpoints are high by construction and are listed for completeness only. ESL = Earth-specific transformer layer.

Stage	3h vs. 4h	3h vs. 5h	3h vs. 6h	4h vs. 5h	4h vs. 6h	5h vs. 6h
Patch embedding	1.000	1.000	1.000	1.000	1.000	1.000
ESL-1	1.000	1.000	1.000	1.000	1.000	1.000
Downsample	1.000	1.000	1.000	1.000	1.000	1.000
ESL-2	0.994	0.984	0.976	0.997	0.993	0.999
ESL-3	0.976	0.951	**0.929**	0.995	0.986	0.998
Upsample	0.999	0.997	0.992	0.999	0.996	0.999
ESL-4	0.999	0.994	0.983	0.998	0.990	0.997
Patch recovery	1.000	1.000	1.000	1.000	1.000	1.000

Table 23: Single-step forecasts of the fine-tuned Pangu-Weather family at the nominal, uncalibrated $\lambda = 1/3$ and $2/3$ (RMSE, mean $\pm$ std over $n = 32$ initializations of 2022; each model at its own lead time). The blends deliver usable forecasts at horizons Pangu-Weather does not ship; at the calibrated $\lambda^\star$ of Table 21 their Z500 RMSE improves to 30.6 and 28.9. For reference, the official 3h model reaches Z500 RMSE 13.14 at 3 h, so the light-budget 3h anchor is about $1.7\times$ worse than the official one; the official 3h model chained autoregressively reaches only multiples of 3 h, so 4 h and 5 h are not reachable from the official checkpoints at all.

Model	Lead	U10 (m/s)	T2m (K)	Z500 (m^2/s^2)	T850 (K)
3h anchor (fine-tuned)	3 h	0.4708 ± 0.0021	0.5609 ± 0.0042	22.23 ± 0.26	0.3903 ± 0.0015
Blend $\lambda = 1/3$	4 h	0.5983 ± 0.0144	0.7222 ± 0.0152	34.61 ± 0.92	0.4956 ± 0.0116
Blend $\lambda = 2/3$	5 h	0.6143 ± 0.0129	0.6861 ± 0.0058	34.78 ± 0.82	0.5079 ± 0.0097
6h model (official)	6 h	0.6082 ± 0.0127	0.6335 ± 0.0049	27.27 ± 0.80	0.5093 ± 0.0090

F.3 GRAPHCAST

Table 24: Mean diagonal CKA per GraphCast stage between the 3h anchor, the 6h anchor and the interpolants at $\lambda = 0.3$ ("4h") and 0.6 ("5h"). The informative column is 3h vs. 6h: the encoder is identical (1.000) and the similarity dips in the processor (0.940) and decoder (0.930), the same mid-network concentration of the horizon adaptation as on Aurora and Pangu-Weather.

Stage	3h vs. 4h	3h vs. 5h	3h vs. 6h	4h vs. 5h	4h vs. 6h	5h vs. 6h
Encoder	1.000	1.000	1.000	1.000	1.000	1.000
Processor	0.996	0.982	**0.940**	0.995	0.965	0.987
Decoder	0.990	0.966	**0.930**	0.993	0.971	0.992

Table 25: Fixed-target sweeps between the GraphCast 3h and 6h anchors at $T^\star = 4\,\mathrm{h}$ and $T^\star = 5\,\mathrm{h}$. $\varepsilon(\lambda) < 0$ at every interior point of both sweeps; the interior minima are 23.24 at $\lambda^\star = 0.30$ (37% below the better endpoint, flat over $[0.25, 0.35]$) and 29.28 at $\lambda^\star = 0.60$ (38% below, flat over $[0.55, 0.65]$).

	$T^\star = 4\,\mathrm{h}$			$T^\star = 5\,\mathrm{h}$		
λ	$L(W_\lambda)$	$L_{\mathrm{lin}}(\lambda)$	$\varepsilon(\lambda)$	$L(W_\lambda)$	$L_{\mathrm{lin}}(\lambda)$	$\varepsilon(\lambda)$
0.000	36.913	36.913	0.000	68.717	68.717	0.000
0.050	33.051	38.441	−5.390	63.715	67.658	−3.943
0.100	29.602	39.970	−10.368	58.906	66.599	−7.692
0.150	26.667	41.498	−14.832	54.168	65.539	−11.371
0.200	24.578	43.027	−18.449	49.854	64.480	−14.626
0.250	23.422	44.556	−21.134	45.726	63.421	−17.695
0.300	**23.244**	46.084	−22.840	41.832	62.362	−20.530
0.350	24.096	47.613	−23.517	38.267	61.303	−23.036
0.400	25.836	49.141	−23.306	35.263	60.244	−24.980
0.450	28.163	50.670	−22.507	32.698	59.184	−26.486
0.500	31.045	52.199	−21.153	30.819	58.125	−27.306
0.550	34.260	53.727	−19.467	29.578	57.066	−27.488
0.600	37.773	55.256	−17.483	**29.279**	56.007	−26.728
0.650	41.430	56.784	−15.354	29.660	54.948	−25.287
0.700	45.132	58.313	−13.181	30.865	53.889	−23.023
0.750	48.945	59.842	−10.897	32.760	52.829	−20.070
0.800	52.699	61.370	−8.672	35.127	51.770	−16.643
0.850	56.438	62.899	−6.461	37.871	50.711	−12.840
0.900	60.213	64.427	−4.215	40.891	49.652	−8.761
0.950	63.841	65.956	−2.115	44.122	48.593	−4.471
1.000	67.485	67.485	0.000	47.534	47.534	0.000

Table 26: Single-step forecasts of the GraphCast family at the calibrated $\lambda^\star$ of Table 25 (latitude-weighted RMSE, mean $\pm$ std over every 6-hourly initialization of 2022, $N = 1460$; each model at its own lead time). The interpolants are usable but, with light-budget anchors, not competitive with a well-tuned native checkpoint: the 4h interpolant's T2m exceeds the 6h anchor's, and its Z500 lies above both anchors' native-lead values, whereas T850 and U10 sit between them. The Z500 values of the interpolants differ from the sweep minima of Table 25 because the sweeps pin the input stencil to 3 h on the 64-initialization subset, whereas this table uses each horizon's own stencil on the full year.

Model	Lead	U10 (m/s)	T2m (K)	Z500 ($\mathrm{m}^2/\mathrm{s}^2$)	T850 (K)
3h anchor (fine-tuned)	3 h	0.3878 ± 0.0187	0.5382 ± 0.0266	23.92 ± 1.27	0.3056 ± 0.0136
LWI $\lambda = 0.3$	4 h	0.4925 ± 0.0996	0.7018 ± 0.0486	28.32 ± 5.25	0.3865 ± 0.0786
LWI $\lambda = 0.6$	5 h	0.5163 ± 0.0882	0.6200 ± 0.0339	26.16 ± 4.86	0.3948 ± 0.0673
6h anchor (fine-tuned)	6 h	0.5050 ± 0.0905	0.5925 ± 0.0328	22.39 ± 4.76	0.3975 ± 0.0639

G ROLLOUT OF INTERPOLATED CHECKPOINTS

Table 27: Autoregressive rollout of the Aurora interpolants $\mathcal{W}_{1/3}$ (4h) and $\mathcal{W}_{2/3}$ (5h), each applied to its own output for two to four steps, evaluated on every 6-hourly initialization of 2022 (RMSE, mean $\pm$ std; Z500 in m^2/s^2, T2m in K). For reference, the native 6h checkpoint rolled out on the 64-initialization subset reaches Z500 25.21 ± 0.94 and T2m 0.51 ± 0.03 at $+12$ h, and 31.95 ± 3.41 / 0.59 ± 0.03 at $+18$ h (Table 5); the interpolants trail it by a factor of about 1.6 on Z500 and 1.3 on T2m at comparable horizons.

Interpolant	Steps	Horizon	Z500	T2m
$\mathcal{W}_{1/3}$ (4h)	2	$+8$ h	32.10 ± 3.16	0.68 ± 0.06
	3	$+12$ h	41.39 ± 2.99	0.68 ± 0.04
	4	$+16$ h	52.02 ± 4.23	0.84 ± 0.07
$\mathcal{W}_{2/3}$ (5h)	2	$+10$ h	35.52 ± 2.64	0.65 ± 0.06
	3	$+15$ h	41.48 ± 3.51	0.74 ± 0.09
	4	$+20$ h	51.78 ± 4.35	0.81 ± 0.08

H INTERPOLATION ACROSS THE 1H–3H GAP

A natural question is whether interpolation also works between $\mathcal{W}_{1h}$ and $\mathcal{W}_{3h}$ to reach 2 h. We checked this across four input-spacing and time-conditioning combinations (Table 28), against the autoregressive 1h + 1h chain of $\mathcal{W}_{1h}$ and against a natively fine-tuned 2h model trained with the recipe of Appendix A.4. The best interpolated configuration is $1.7\times$ worse than the chain on Z500 and $1.8\times$ worse than the native 2h model, mirroring the CKA collapse between $\mathcal{W}_{1h}$ and the other two checkpoints.

Table 28: 1h–3h interpolation to 2 h (RMSE, mean $\pm$ std; the LWI and chain rows on the 64-initialization subset of 2022, the same sample and script as Table 5; the native 2h row on every 6-hourly initialization, $N = 1460$). The four LWI rows vary input spacing (1 h vs. 2 h) and time-conditioning embedding (1 h vs. 2 h). The reference rows are the autoregressive 1h + 1h chain of the native $\mathcal{W}_{1h}$ model and a natively fine-tuned 2h model. The best LWI configuration (spacing/conditioning 2/1) is $\sim 1.7\times$ worse than the chain on Z500 RMSE, illustrating the loss of connectivity once the endpoints leave the basin.

Setup	spacing/time cond. (h)	Z500	T850	T2m	U10
LWI	1/1	24.78 ± 1.11	0.251 ± 0.007	0.522 ± 0.061	0.307 ± 0.010
LWI	1/2	26.38 ± 1.50	0.284 ± 0.007	0.646 ± 0.021	0.424 ± 0.007
LWI	2/1	20.71 ± 0.91	0.23 ± 0.01	0.519 ± 0.049	0.291 ± 0.011
LWI	2/2	23.00 ± 1.30	0.276 ± 0.008	0.640 ± 0.031	0.414 ± 0.008
Chain 1h + 1h	1/1	12.12 ± 0.73	0.213 ± 0.006	0.48 ± 0.04	0.26 ± 0.012
Native 2h (fine-tuned)	2/2	$\mathbf{11.73 \pm 0.72}$	$\mathbf{0.21 \pm 0.01}$	$\mathbf{0.43 \pm 0.02}$	$\mathbf{0.24 \pm 0.02}$

I FORECAST VISUALIZATIONS

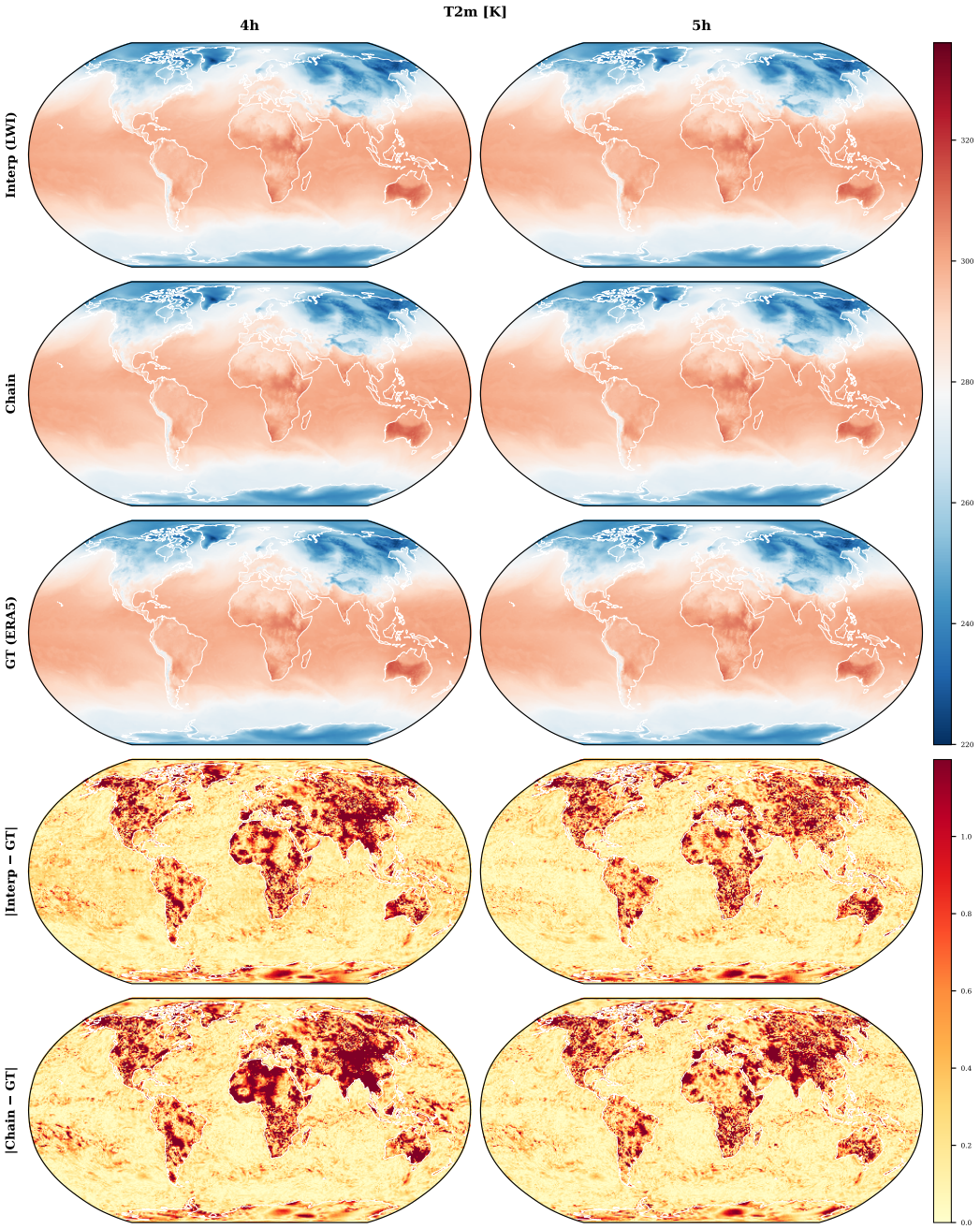


Figure 6: Forecast maps for 2 m temperature (T2m) at 4 h and 5 h on one 2022 initialization: ERA5 target, LWI, the autoregressive chain and the grid-space interpolants, with the corresponding error fields.

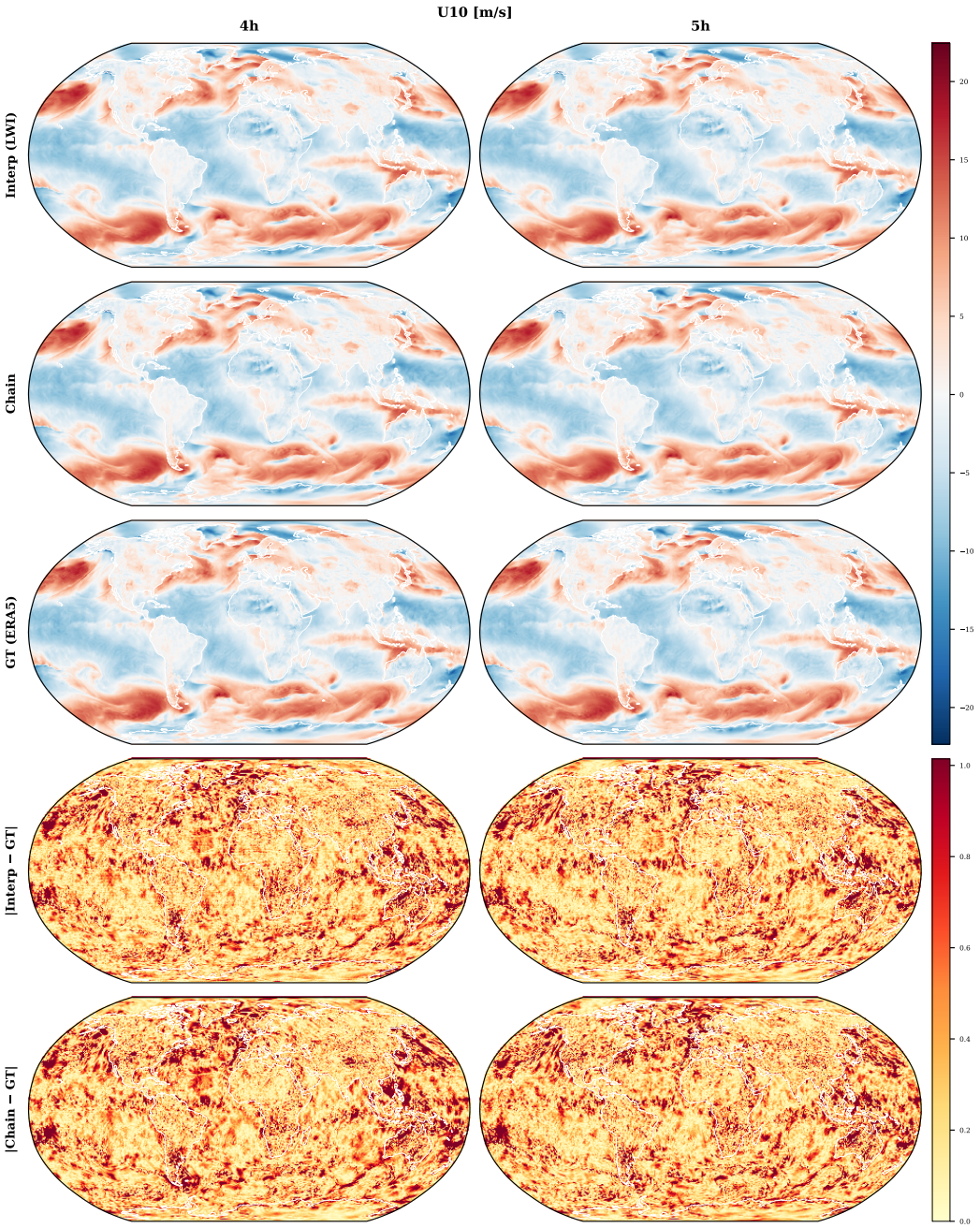


Figure 7: Forecast maps for 10 m zonal wind (U10) at 4 h and 5 h on one 2022 initialization: ERA5 target, LWI, the autoregressive chain and the grid-space interpolants, with the corresponding error fields.

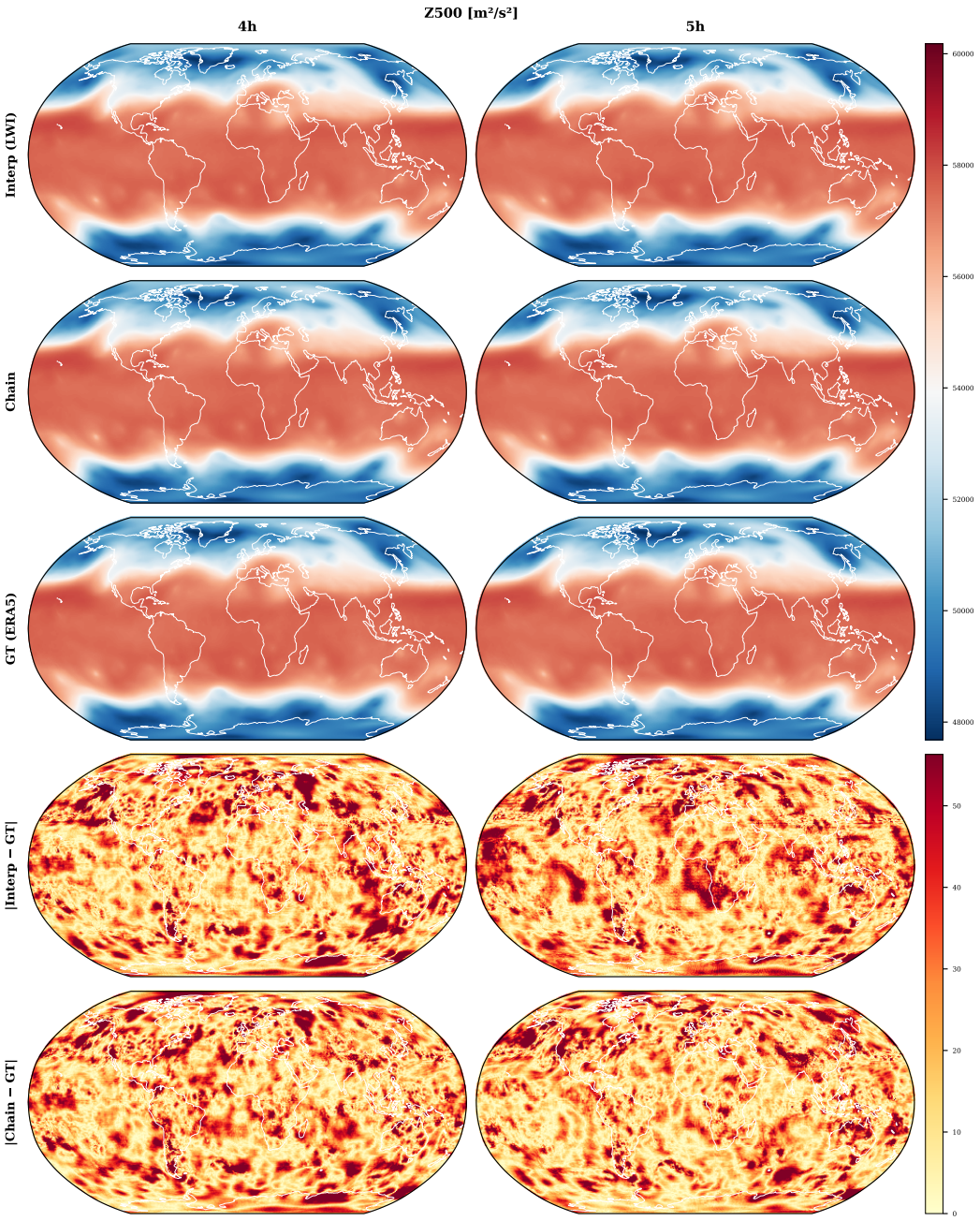


Figure 8: Forecast maps for 500 hPa geopotential (Z500) at 4 h and 5 h on one 2022 initialization: ERA5 target, LWI, the autoregressive chain and the grid-space interpolants, with the corresponding error fields.

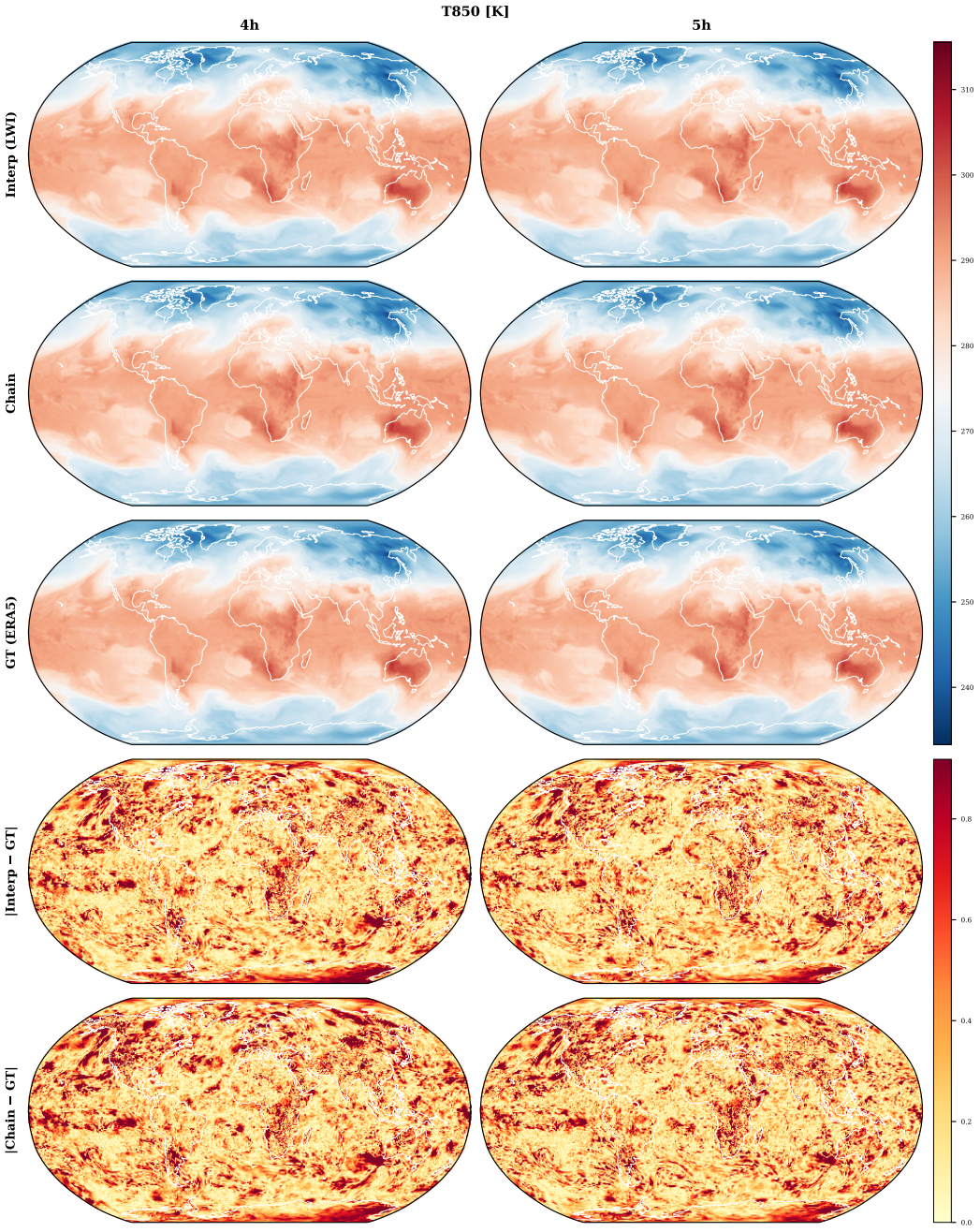


Figure 9: Forecast maps for 850 hPa temperature (T850) at 4 h and 5 h on one 2022 initialization: ERA5 target, LWI, the autoregressive chain and the grid-space interpolants, with the corresponding error fields.

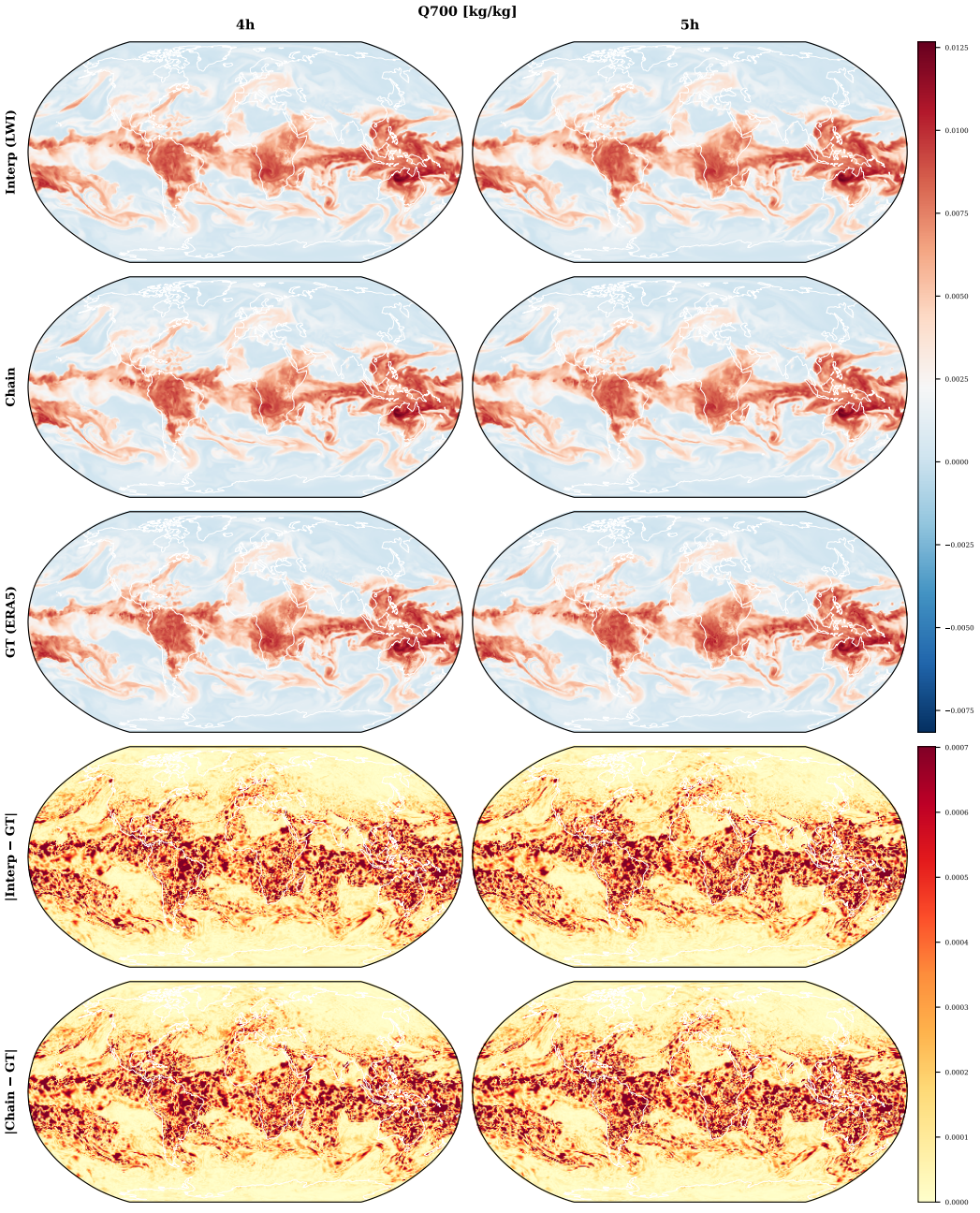


Figure 10: Forecast maps for 700 hPa specific humidity (Q700) at 4 h and 5 h on one 2022 initialization: ERA5 target, LWI, the autoregressive chain and the grid-space interpolants, with the corresponding error fields.