A Path to Sub-Km Regional Emulation with the NOAA HRRRCast: Lateral Height-Dependent Boundary Forcing and a Resolution-Invariant Nesting Methodology

Fnu Sirisha and Milton Halem

University of Maryland, Baltimore County (UMBC), Baltimore, Maryland, USA.

Corresponding author: Milton Halem (halem@umbc.edu).

Preprint, November 2, 2026.

Abstract

Two non- retraining methods move HRRRCast toward nested sub-kilometer forecasting, (i) Height-dependent lateral relaxation toward GFS cuts 48-hour forecast errors of the 6 km HRRRCast diffusion model by 14–43% in a summer and a winter case, and (ii) coupling a nest to a CONUS HRRRCast host, with the same relaxation at the nest's edges, lets the same 6 km model run on 3 km sub-domains with lower error than its own CONUS forecast over the same area.

HRRRCast advances its state at a 6-hour cadence and sees GFS only as a full-domain conditioning at four pressure levels. Its errors grow fastest near the domain edges: in summer along the northern boundary, in winter across the whole domain and most at the upwind (western) edge. We relax the model state toward regridded GFS in edge zones whose width at each level equals the distance air travels in 6 hours at a climatological wind speed (22 cells at 1000 hPa to 180 cells at 200 hPa on the 6 km grid). The zones are applied at every 6-hour boundary, on all 20 levels and the 10 m and 80 m winds.

At 48 hours, a whole-domain RMSE is reduced 29–39% for 500 hPa height and 14–20% for mean sea-level pressure, which is never relaxed directly. Edge-band errors are reduced 60–82%; interior gains are smaller (3–4% in summer, 7–14% in winter). Relaxing only the four GFS conditioning levels makes every unrelaxed level worse (up to 2.6× baseline in edge bands). Relaxing only at inflow points, the physics-model convention, removes most of the benefit: the network's edge errors arise at outflow edges too.

We also find that HRRRCast's GFS inputs pair earth-relative GFS winds with grid-relative HRRR winds on the Lambert grid, a rotation of up to 23° at the east and west edges. This is still present in the current public code. The network appears to have learned to compensate, which hides the error from standard verification.

Run on native 3 km input over four nested sub-domains, the 6 km HRRRCast matches its own 6 km skill on the same domain (error ratios 0.7–1.2 to 48 hours), supporting near-resolution invariance. Small stand-alone nests whose western edge lies over land collapse within one 6-hour step at either resolution. We trace this to the normalization of the nested OROG network's global max pooling: taking the max-pooled values of a full-CONUS HRRRCast, run in parallel with the nested HRRRCast run, makes the NE nest's 6-hour 500 hPa height error reduce from 268 m to 19 m, matching the CONUS forecast. This coupled CONUS HRRRCast -in-HRRRCast nesting. through these max pools at every 6-hour step, avoids retraining the model. The NE and SE nests at both 3 km and 6 km stay near CONUS skill to 48

hours (F48 500 hPa height error 0.90–1.28 of the CONUS run, against 2.4–3.1× for the stand-alone nests), with 3 km and 6 km nests within 0.96–1.28 of each other. Adding the GFS relaxation zones at the coupled nests' edges lowers the F48 500 hPa height error to 0.54–0.86 of the CONUS run (3 km nests; 0.58–0.90 at 6 km). Relaxing the edges toward the CONUS HRRRCast host instead gains only about 10%, since it adds no information the host lacks.

These results outline a path to 1 km nests built from GSL's native 3 km HRRRCast, by the same steps: initialize the nest from a 1 km analysis, normalize static fields with the model's fixed training statistics, couple the nest to a CONUS host through the global max pools (or retrain on random sub-domain crops), and relax toward the driving model in zones sized to 6-hour advection. Resolution invariance from 3 km to 1 km is the hypothesis we test next.

Keywords: AI weather prediction; HRRRCast; limited-area modeling; lateral boundary conditions; nesting; resolution invariance; channel attention (CBAM); diffusion models.

1. Introduction

AI weather models (GraphCast, Lam et al. 2023; Pangu-Weather, Bi et al. 2023; FourCastNet, Pathak et al. 2022; Aurora, Bodnar et al. 2025; Prithvi WxC, Schmude et al. 2024) now rival physics models at global scale. Recently, the 6km and now 3km regional AI model HRRRCast (NOAA GSL; Abdi et al. 2026a) brings an operational approach of convection-allowing resolution emulation over CONUS. A regional model must take in information from outside its domain. Physics Limited-Area Models (LAM) do this with lateral boundary conditions and relaxation zones (Davies 1976; Warner et al. 1997), as in NCAR's WRF model (Skamarock et al. 2019) and NOAA's FV3 limited-area model (Black et al. 2021), applied every time step. An AI regional model needs a lateral boundary treatment as much as a physics model does, and HRRRCast's current one is weak at long lead times. (Hamer et. al.,) explored the impact of various relaxation zones provided by a global AI model on the ECMWF CERRA reanalysis. HRRRCast instead conditions each 6-hour forecast step on GFS fields at four pressure levels over the whole domain.

Our longer-term goal is sub-km AI nested forecasts (1 km, 300 m) for wildfires, flash floods and severe-storm digital twins, where physics models are too slow for real-time ensembles. A nested AI model needs boundary forcing that works on any sub-domain and does not rely on a coarse global model forecasts across the whole interior.

This paper makes six contributions:

1. A diagnosis of where HRRRCast's 6 km forecast error grows, by distance from each domain edge.
2. A relaxation-zone method sized to the 6-hour cadence, and its effect on two contrasting cases.
3. Ablations: partial-level relaxation, no surface relaxation, a narrower top zone, and inflow-only (characteristic) relaxation.
4. A wind-frame mismatch between HRRRCast's GFS inputs and its HRRR state, which the HRRRCast network training appears to have learned to compensate the internal and external wind state representation.
5. A near resolution-invariance test: the 6 km checkpoint run on native 3 km input over four sub-domains matches its 6 km skill on the same domain, once a static-field normalization error that distorts any sub-domain is corrected (Sections 7 and 10).

6. A diagnosis of why small stand-alone nests collapse (the network's global max pooling) and an HRRRCast-in-HRRRCast coupling that removes the collapse to 48 hours without retraining; with the GFS relaxation zones of contribution 2 added at the nest edges, 3 km nests beat the CONUS run over their own area at 48 hours (Section 10).

Together, contributions 2, 5 and 6 give the ingredients of a sub-kilometer nest: boundary forcing sized to the model's cadence, a model that keeps its large-scale skill on a finer grid, and a host coupling that makes small domains usable. Section 11 relates them to GSL's native 3 km HRRRCast, the natural model for a 1 km nest.

2. Model, cases and verification

All experiments are conducted on an ASUS game computer for under $2K and the use of Claude for programming. Hallucination checking was provided by conducting external validation runs on the UMBC CHIP computer and the NAIRR PSC computer. We employ the public 6 km HRRRCast checkpoint `3355614` (NOAA-GSL/HRRRCast-live, Jan 2026), run unchanged except for the boundary condition treatment under test.

- **Grid and state:** 530 × 900 cells at 6 km (the 3 km HRRR grid decimated by 2), Lambert conformal (LoV 262.5°E, standard latitude 38.5°N). The state has 6 variables (U, V, VVEL, TMP, HGT, SPFH) on 20 pressure levels from 200 to 1000 hPa, plus 16 surface fields.
- **GFS conditioning:** All 6 GFS variables at 250, 500, 850 and 1000 hPa, plus 16 surface fields (40 channels), are regridded bilinearly to the full domain from the 0.25° GFS cycle before initialization, using GSL's xESMF-based preprocessing (src/make_bcs.py).
- **Sampling:** DDIM diffusion sampling (Song et al. 2021) with 25 denoising steps, single member (m0). Each 6-hour step costs about 15 minutes on the ASUS CPU.
- **Cases:** summer, initialized 2026-08-10 09Z (weak flow, active northern boundary), and winter, initialized 2026-01-06 09Z (strong westerly jet). Both run to 48 hours as eight chained 6-hour steps (a 6-hour cadence).
- **Verification:** RMSE against the HRRR analysis (Dowell et al. 2022) decimated to the same 6 km grid (pixel-exact alignment). Scored fields: MSLP, 10 m U, and HGT, TMP and vector wind at 300, 500 and 900 hPa. At 900 hPa, points where analysis surface pressure is below 900 hPa are excluded.
- **Regions:** the whole domain; bands 0–50 and 50–100 cells in from each edge (50-cell corners excluded); and the interior beyond the 500 hPa zone (distance ≥ 97 cells) and beyond the 200 hPa zone (≥ 180 cells).

Scores for relaxed runs are taken from the model's own 6-hour step *before* that boundary's relaxation, so they measure the forecast, not inserted GFS values.

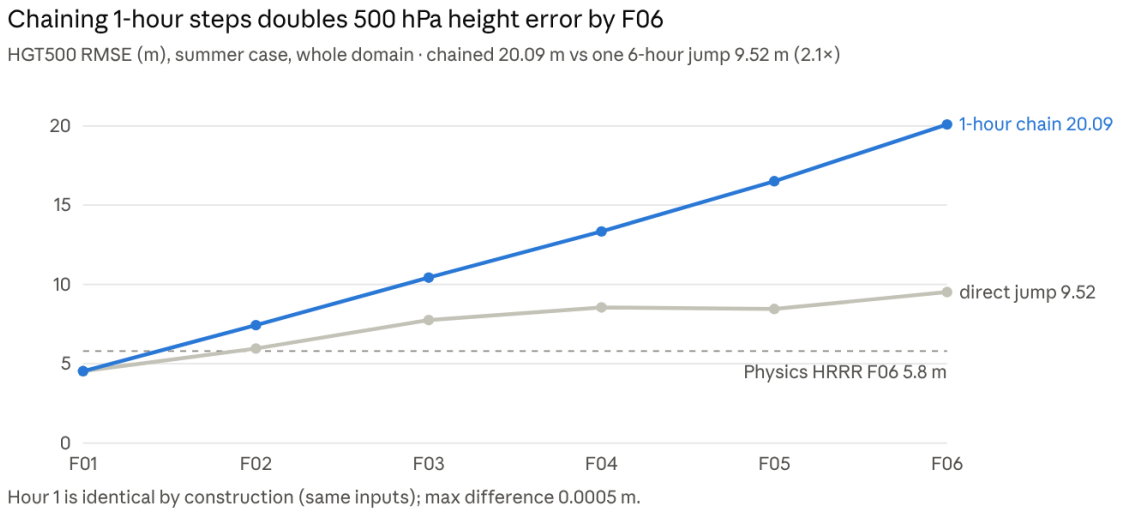
Robustness to grid-scale noise. The chained diffusion forecasts accumulate grid-scale roughness: the mean HGT500 gradient is 2.3× the analysis value at F06 and 17× at F48 (11.5× with relaxation zones). Its amplitude is small, however: smoothing forecast and analysis identically with a Gaussian filter of $\sigma = 3$ cells (about 18 km) lowers RMSE by only 3–5% and leaves every run-to-baseline ratio essentially unchanged. The results below therefore reflect large-scale errors.

3. How GFS enters HRRRCast

GFS reaches the evolving forecast only once per 6 hours, so a physics-style boundary update every time step has no direct equivalent.

For each output hour h, the network takes the state at the last 6-hour point, GFS at hour h, and a lead encoding (h − last 6-hour point)/6, and steps directly to hour h. Hours 1–5 of each block are side outputs; only the hour-6 output starts the next block. GFS at hours 1–5 therefore never affects later forecasts, and in a 48-hour run GFS enters the forecast chain eight times.

A finer cadence would need 1-hour chaining, each step starting from the previous hour's output. We tested this to F06 on the summer case:



HRRRCast 3355614, summer case 2026-08-10 09Z, member 0, vs decimated HRRR analysis

Chained error grows about twice as fast as the direct 6-hour step, and the gap widens with each step. The 1-hour chained F06 is about as wrong as the normal model at F18. Boundary treatment must therefore work at the model's 6-hour cadence.

4. Where HRRRCast's error grows

Without relaxation, the 48-hour error concentrates at the northern edge in summer and is large everywhere, worst upwind, in winter.

Baseline HGT500 RMSE at F48 (m)	Summer (2026-08-10)	Winter (2026-01-06)
Whole domain	79.6	149.0
West edge, 0–50 cells	63	185
North edge, 0–50 cells	176	182
East edge, 0–50 cells	45	100

Baseline HGT500 RMSE at F48 (m)	Summer (2026-08-10)	Winter (2026-01-06)
South edge, 0–50 cells	37	60
Interior (≥ 97 cells from every edge)	63	151

Context: GFS is far more accurate at the synoptic scale. GFS regridded bilinearly to the same 6 km grid (as for the model's own GFS inputs) and verified against the same HRRR analysis decimated to 6 km has 48-hour HGT500 RMSE of 7.3 m (summer) and 11.5 m (winter), against 79.5 m and 149.0 m for this HRRRCast checkpoint, which is conditioned on that same GFS. The gap is 2–3× at F06 and 11–13× at F48; in winter HRRRCast also develops a mean height bias of −67 m by F48 (GFS −6 m). RMSE favors smooth fields and does not measure fine-scale structure: GFS at 0.25° contains no convective-scale detail, which HRRRCast does represent (as the power spectra of Abdi et al. 2026b show), and realistic but slightly misplaced small-scale features are penalized twice. The comparison therefore concerns synoptic-scale accuracy only. At 500 hPa that is what the gap reflects: smoothing forecast and analysis over about 18 km lowers HRRRCast's RMSE by only 3–5% (Section 2), so its error is large-scale drift rather than misplaced detail. The relaxation gains reported below are relative to a model that drifts strongly from its own conditioning; even the best configuration remains 6–9× worse than GFS at 500 hPa at F48. GSL's current 3 km model is much stronger (single member about 28.5 m at F48 on another date), so the results should be confirmed with it.

- **Summer:** through F18 the edge bands are no worse than the interior, and the western (upwind) strip is the best region. From F12 the northern 50 rows (about 49–52.6°N) diverge. The AI-to-physics error ratio there rises from 2.5 at F06 to 6.3 at F18, against 3.3–4 in the interior and 1.1–3 at the southern edge. The excess peaks 10–50 rows in and widens from about 50 rows at F24 to 75–100 rows by F48.
- **Winter:** error is about twice the summer level and spread through the interior. The western edge, where the Pacific jet enters, is worst; MSLP at the eastern edge reaches 12.2 hPa as a deepening system leaves the domain.

In both cases, a 6-hour step must account for air that crosses roughly 130 km near the surface and over 1,000 km at jet level: information the network cannot see from inside the domain.

5. Relaxation zone design

Each level's zone is as wide as the distance air travels in one 6-hour step at a climatological wind speed.

Anchor speeds are 6, 11, 27 and 50 m/s at 1000, 850, 500 and 200 hPa, interpolated linearly in ln p for the other levels. Width in cells is the speed times 21,600 s divided by 6 km:

Level (hPa)	Wind (m/s)	6-h distance (km)	Zone width (cells)
1000 and surface	6	130	22
850	11	238	40
500	27	583	97

Level (hPa)	Wind (m/s)	6-h distance (km)	Zone width (cells)
300	(interpolated)	860	143
200	50	1,080	180

At each 6-hour boundary, after the step and before the next, the state is blended toward GFS regridded to the 6 km grid with the same bilinear method the model's own GFS inputs use:

$$x \leftarrow x + w(d)\,(x_{\mathrm{GFS}} - x), \qquad w(d) = \max(0,\, 1 - d/W_\ell)$$

where d is the distance in cells from the nearest edge and Wℓ the zone width at level ℓ. The weight is 1 at the edge and falls linearly to 0 at the zone's inner edge; everything beyond is untouched. The blend is done in normalized units, which equals blending physical values because normalization is linear.

- **Relaxed:** UGRD, VGRD, TMP, HGT, SPFH on all 20 levels (GFS has every level except 825 and 875 hPa, which are interpolated in ln p), plus the 10 m and 80 m winds with the 1000 hPa width.
- **Not relaxed:** VVEL (25 km GFS vertical motion is not comparable to HRRR's), surface pressure, 2 m temperature and dewpoint (GFS-vs-HRRR terrain differs by hundreds of meters), MSLP (different reduction), and diagnostic fields.
- **Unchanged:** the model's own 40-channel GFS conditioning, except in the experiments that remove it (Section 9).

6. Results: relaxation all around the edges

Relaxing all 20 levels and the surface winds (run A) reduces 48-hour error in every scored field in both seasons; the gains come mostly from the edges.

Run A’ (wind targets rotated to the HRRR grid, Section 7) gives the same whole-domain and interior scores as run A to within about 1% in both seasons, so the run-A numbers below stand. A’ differs at the edges: it is better at the east (outflow) edge in both seasons, and in winter slightly worse for 900 hPa wind and temperature at the strong-inflow west edge.

- **Edges:** HGT500 at F48 are reduced by 60–82% in the 0–50 cell bands of every edge in both cases (summer north 176 → 32 m; winter west 185 → 33 m). The summer northern-edge blow-up disappears.
- **Interior:** beyond the 500 hPa zone (≥ 97 cells), gains appear late and stay small: HGT500 −4% in summer at F48, −10% in winter, where they already reach −4% by F24. Stronger winter flow carries boundary information inward faster.
- **Weak spot:** MSLP in the southern 0–50 band is 17–26% worse than baseline at F24–F36 in both cases, returning to neutral by F48.

Two variants isolate the design choices (ratio to baseline, whole domain, F48):

Variant	Summer HGT500	Summer MSLP	Summer U10	Winter HGT500	Winter MSLP	Winter U10
A: all 20 levels + surface winds	0.61	0.86	0.78	0.71	0.80	0.83
B: only 1000/850/500/200 hPa + surface winds	0.80	1.17	1.04	—	—	—
C: top zone 120 cells, no surface winds	0.62	0.89	0.91	0.71	0.84	0.92

- **B, partial levels:** relaxed levels improve, but every unrelaxed level and MSLP get worse: HGT300 +20%, TMP900 +42% at F48, up to 2.6× baseline in edge bands. Mixed columns (GFS-like on four levels, model-like between) are not reconciled by the network. The zones must cover the full column.
- **C, narrower top, no surface winds:** the 120-cell top zone costs little at 500 hPa and a few percent at 300 hPa. Dropping the 10 m and 80 m winds makes U10 worse than baseline in outer edge bands: summer west, east and south 1.35×, 1.21×, 1.46×; winter west and south 1.09× and 1.15×, against 0.44× and 0.61× with A (all F48). Relaxing 1000 hPa does not substitute for the near-surface wind channels.

7. A hidden wind-frame mismatch in HRRRCast's GFS inputs

HRRRCast conditions on GFS winds relative to true north and east while predicting HRRR winds relative to its Lambert grid; the two differ by up to 23° at the domain's east and west edges.

HRRR GRIB winds are grid-relative (`uvRelativeToGrid = 1`); GFS 0.25° winds are earth-relative (`= 0`). The preprocessing in `src/make_bcs.py` regrids GFS u and v with xESMF without rotating them. xESMF interpolates u and v as separate scalar fields, so each wind lands at the right grid point but keeps its earth-relative direction; the earth-to-grid rotation must be applied as a separate step after regridding. This is unchanged in the current public master (`b81da9a`, 8 Oct 2026, which includes the 3 km network); the only rotation code in the repository is in `src/diagnostics.py`, for derived diagnostics. The local angle and the earth-to-grid rotation are:

$$\alpha = \sin(38.5^\circ)\,(\lambda - 262.5^\circ), \qquad u_g = u_e\cos\alpha - v_e\sin\alpha, \qquad v_g = u_e\sin\alpha + v_e\cos\alpha$$

The angle is about 0° at 97.5°W and reaches ±22.8° at the west and east edges. Checked against the HRRR analysis at the same time, rotated GFS winds match far better at the edges (winter case, vector RMS difference, 100 columns from each edge):

Level	West, unrotated	West, rotated	East, unrotated	East, rotated	Mid-domain
300 hPa	12.7 m/s	3.6 m/s	10.7 m/s	2.3 m/s	2.7–2.8 m/s
500 hPa	7.8 m/s	3.5 m/s	6.4 m/s	1.7 m/s	2.8 m/s
900 hPa	3.6 m/s	2.0 m/s	3.5 m/s	2.3 m/s	3.0 m/s

The network has likely learned to compensate. The rotation is the same fixed pattern in every training sample, so a network trained on it can learn to reinterpret GFS winds by location; its outputs are verified against HRRR in HRRR's own frame, so training loss and verification would not reveal the mismatch. This resembles shortcut learning (Geirhos et al. 2020): the correction becomes a hidden dependency that shows itself only when the GFS–state relationship changes.

Runs A, B and C relaxed the edge winds toward unrotated targets. Repeating A with rotated targets (A', winter) leaves the 48-hour whole-domain and interior scores unchanged (HGT500 0.70 vs 0.71 of baseline) but moves where the skill comes from:

- **East (outflow) edge:** better with correct winds at every level from F12 (300 hPa wind 0–50 band 0.24 vs 0.28; 900 hPa temperature 0.47 vs 0.63 at F48).
- **West (inflow) edge:** upper heights slightly better by F48, but 900 hPa wind and temperature are worse throughout (wind 0.46 vs 0.37). Low-level inflow is where the network appears most tuned to the unrotated GFS winds.

In the summer case, with weak flow, A' is equal to or better than A everywhere, including the west edge: at F48, 300 hPa wind in the 0–50 bands is 0.37 vs 0.42 (west) and 0.40 vs 0.56 (east), HGT500 west 0.34 vs 0.39, 900 hPa temperature west 0.93 vs 1.01; whole-domain scores improve by about 1%. The low-level penalty is therefore specific to strong inflow, where the network relies most on its unrotated GFS winds.

A' is the reference for all later runs.

Rotating the model's own GFS wind inputs to the HRRR frame ("rotforce", winter, no relaxation) gives no net gain without retraining, and its effect is confined to the edges:

- **Whole domain:** 0.98–1.04 of baseline at F24 and 0.96–1.02 at F48 across all nine scored fields.
- **Interior (≥ 97 cells):** 0.98–1.03 throughout, although the rotation there still reaches about ±18°. The network leans on GFS wind direction mainly where air enters, not across the interior.
- **West (inflow) edge:** near-surface and upper winds are worse early (U10 0–50 band 1.37 at F18, 1.22 at F24; 300 hPa wind 1.17 at F12), while low-level mass fields are better by F48 (MSLP 0.69, HGT900 0.65).
- **East (outflow) edge:** 300 hPa wind better from F12 (0.81–0.83 at F18–F24); HGT500 worse (1.12–1.19); 900 hPa temperature and wind better by F48 (0.87, 0.88).

The mixed, edge-local response is what a network that learned the unrotated winds would show. Fixing the mismatch properly means rotating GFS winds in `make_bcs.py` and retraining, which we recommend to GSL.

A second preprocessing issue: static fields normalized per domain. The two static inputs, land mask (LAND) and orography (OROG), are missing from the checkpoint's `normalize-stats.nc`. `src/make_ics.py` then standardizes them with the mean and standard deviation of whatever domain it is given, a fallback still present in public master (`b81da9a`). On the full 6 km CONUS grid this reproduces what the network saw in training; on any sub-domain it does not.

Domain	OROG mean (m)	OROG std (m)	Highest peak, normalized
6 km CONUS (training-like)	476	658	5.35 (Rockies)
SW crop, native 3 km	735	805	4.05 (same peak)
NE crop, native 3 km	254	171	6.82 (1419 m, Appalachians)

On the NE crop the network sees Appalachian terrain taller than the Rockies and sea level far too low. Its winter forecast collapsed cold within one 6-hour step: HGT500 bias −243 m and 900 hPa temperature bias −13 K at F06, 6 to 18 times the 6 km error over the same area. Re-normalizing LAND and OROG with the CONUS statistics corrects the inputs; the SW crop's highest peak then normalizes to 5.35, as on CONUS.

Our earlier native-3 km crop tests, which suggested the 6 km checkpoint does not transfer to 3 km input, used the distorted statics and must be repeated. Corrected runs for four nests, with and without A' zones, are reported in Section 10. We recommend that GSL store LAND and OROG statistics in `normalize-stats.nc` so that any nested or cropped domain uses the training values.

8. Inflow-only relaxation

Relaxing only where air enters the domain, the physics-model convention, removes most of the benefit of relaxation in the winter case.

The inflow gate is deliberately simple. At each 6-hour boundary and for each level:

1. Each zone point is assigned to its nearest edge.
2. The GFS wind component normal to that edge is read at a single point: where the straight grid line through the zone point, perpendicular to the edge, meets the edge (west u, east −u, south v, north −v; positive = inflow).
3. That one value sets a gate that ramps from 0 at 2 m/s outflow through 0.5 for tangential flow to 1 at 2 m/s inflow.
4. The gate is applied unchanged along the whole perpendicular line into the domain, multiplied by the linear distance ramp of Section 5.

The gate therefore uses the wind at one instant (the end of the step), not along a 6-hour path; the relaxed region follows grid lines perpendicular to the edge, so oblique inflow is treated as if it entered straight in; and the zone width remains the fixed climatological 6-hour distance, not the distance actually travelled that day. All variables at a level share that level's gate; the surface winds use the 10 m wind. At F06 the mean edge gates at 200 hPa were west 0.67, east 0.00, south 0.72, north 0.67, as expected for a westerly jet.

The edge-normal component is used only to set the gate. The relaxation itself always acts on the full wind vector: u and v are both relaxed toward GFS with the same weight at each point, in this run and in all others. Because the grid edges run along the Lambert grid's x and y axes and the winds are grid-

relative, the normal component is exactly u at the west and east edges and v at the south and north edges.

Winter, F48, ratio to baseline	A′ (all edge points)	Inflow-only
Whole domain: HGT500 / MSLP / U10	0.70 / 0.80 / 0.83	0.81 / 0.91 / 0.90
Whole domain: HGT300 / wind 300 hPa	0.71 / 0.74	0.82 / 0.87
Interior (≥ 97 cells): HGT500 / HGT300	0.90 / 0.86	0.94 / 0.93
HGT500, east 0–50 (outflow)	0.33	0.84
HGT500, north / south / west 0–50	0.28 / 0.37 / 0.16	0.58 / 0.55 / 0.22
MSLP, south 0–50	1.03	0.95

The eastern (outflow) edge loses nearly all its gain, and the partially gated edges lose part of theirs. The only improvements are the southern-edge MSLP weak spot (F24: 1.05 vs 1.23) and western 900 hPa temperature (0.48 vs 0.55).

In a physics model, specifying outflow boundaries over-constrains the equations and reflects waves. HRRRCast's edge errors are different: the network has no information beyond the domain and degrades at every edge regardless of flow direction. Relaxation toward GFS repairs that, so it belongs all around the boundary.

Proposed: trajectory-based zones. A closer implementation of the 6-hour advection idea would let the flow, not the grid, set each zone's shape and depth:

1. For every point near the boundary, compute a 6-hour back-trajectory from GFS winds, interpolated bilinearly in space and linearly in time between GFS at t, t + 3 h and t + 6 h.
2. If the parcel was outside the domain at any time in the past 6 hours, that air came from where only GFS knows the state: relax toward GFS there, weighted by the fraction of the 6 hours spent outside.
3. The zone then becomes wide and flow-shaped where inflow is strong, follows oblique entry and corners, and shrinks to nothing where flow is outbound or calm.

Because the network's errors arise at every edge regardless of flow direction (above), trajectory zones should be combined with a narrow fixed band all around the boundary: the band repairs the network's edge handling, and the trajectory zone reaches deeper inland where real inflow occurs. The cost is a vectorized trajectory calculation over the zone points, a few seconds per boundary. This tests whether flow-dependent depth beats the fixed climatological widths.

The summer case confirms this in weak flow, where the F06 gates differ sharply from winter (200 hPa west 0.96, but 850 hPa west only 0.25; south 0.81 at 850 hPa with Gulf inflow):

Summer, F48, ratio to baseline	A′ (all edge points)	Inflow-only
Whole domain: HGT500 / MSLP / U10	0.60 / 0.86 / 0.77	0.65 / 0.89 / 0.81

Summer, F48, ratio to baseline	A′ (all edge points)	Inflow-only
Whole domain: HGT300 / wind 300 hPa	0.56 / 0.76	0.66 / 0.84
HGT500, east / north / west 0–50	0.29 / 0.17 / 0.34	0.92 / 0.25 / 0.44
MSLP, south 0–50	1.03	0.89

In both cases the only consistent gain from inflow-only gating is southern-edge MSLP, which ties that weak spot to relaxing outflow and along-edge points on the Gulf and Mexico boundary. A targeted fix, such as gating only the southern edge, is a candidate refinement.

9. Follow-up experiments

Five further runs test the wind rotation and the role of the model's interior GFS conditioning:

Experiment	Case	What it tests	Reported in
rotforce	Winter	Model's own GFS wind inputs rotated to grid-relative, no relaxation	Section 7
Zones only	Winter	A′ zones with the 24 GFS pressure-level input channels set to zero in normalized units, a flat constant carrying no GFS or climatological information (16 surface channels kept)	This section
A′	Summer	Corrected-wind reference for the summer case	Section 7
Inflow-only	Summer	Section 8 test in weak summer flow	Section 8
No pressure-level GFS, no zones	Winter	Cost of removing GFS on its own, to read the zones-only run against	This section

Zones-only result (winter). Removing the GFS pressure-level conditioning everywhere more than doubles the error at first, and the A′ zones then win it back; by F48 the forecast beats the full-GFS baseline for heights, temperatures and MSLP, but not for winds.

How GFS was removed. The 24 GFS pressure-level input channels (6 variables at 250, 500, 850 and 1000 hPa) were set to 0 in the model's normalized units. HRRRCast normalizes each channel with one mean and one standard deviation for the whole domain, so 0 equals that channel's training-data mean: the same constant at every grid point and every hour. No climatological field, analysis or GFS value is inserted; the network receives a flat, information-free input where GFS used to be. The 16 GFS surface channels were kept. The companion run without zones uses the same zeroing.

Zones only, ratio to full-GFS baseline	F06	F12	F24	F36	F48
HGT500, whole domain	2.52	1.94	1.25	0.95	0.80
HGT500, interior (≥ 97 cells)	2.40	2.25	1.48	1.11	1.00

Zones only, ratio to full-GFS baseline	F06	F12	F24	F36	F48
MSLP, whole domain	2.98	3.26	2.30	1.56	1.01
U10, whole domain	1.35	1.39	1.20	1.17	1.15

At F48, A' (zones plus GFS conditioning) remains better: HGT500 0.70, MSLP 0.80, U10 0.83, 300 hPa wind 0.74 against zones-only 0.80, 1.01, 1.15 and 0.92. Upper-level heights recover fastest (HGT300 0.92 by F24, 0.71 at F48, equal to A'), helped by the wide upper zones and the winter jet. Interior winds stay worse (300 hPa wind 1.19 at F48), consistent with the network leaning most on its interior GFS winds. The early penalty reflects a checkpoint never trained without interior GFS; a model trained with boundary-only forcing should avoid much of it.

Companion run (no pressure-level GFS, no zones) — a correction to the reading above. Much of the apparent recovery happens without zones too, so it is not the zones' doing: whole-domain HGT500 goes 2.52, 2.22, 1.60, 1.18, 0.93 of baseline at F06–F48 with neither GFS nor zones. The zones' contribution is at the edges and in domain-wide upper-air fields (F48 whole-domain HGT500 0.81 with zones vs 0.93 without; edge bands 0.30 vs 0.85; 300 hPa wind 0.93 vs 1.19). In the interior the two runs are nearly equal (HGT500 1.00 vs 0.94, MSLP 1.03 vs 1.02), with only interior winds helped by the zones. That removing GFS costs little by F48 is consistent with the context in Section 4: by day two this checkpoint no longer tracks its GFS conditioning closely.

The zones-only pair tests the structural limitation of HRRRCast's design: conditioning on GFS everywhere makes it partly a GFS downscaler, which a nested sub-km model cannot rely on. The network was trained with full-domain GFS at four levels and to the best of our knowledge, never with relaxed boundaries at all levels, so both runs are out of its training distribution.

Future work. Nested sub-domains in summer, and nests with zones only (GFS surface forcing kept); a third case, ideally spring or a strong summer jet event; trajectory-based zones (Section 8), a fixed narrow band all around plus 6-hour back-trajectory relaxation, in both cases; ensembles of several members; and the nesting tests of Section 10 repeated with GSL's native 3 km HRRRCast, toward 1 km nests (Section 11).

10. Nested domains

The 6 km checkpoint is close to resolution-invariant: run on native 3 km input it matches its 6 km skill. It runs stand-alone, however, only on domains who's western (inflow) edge lies over the Pacific, or that are nearly as large as CONUS; even the Pacific-bounded NW nest drifts to about twice the CONUS error by 48 hours.

Setup. Nests are 530 × 900-cell windows of the native 3 km HRRR grid (no decimation), the same array size as the 6 km CONUS grid: SW (22.6–40.1°N, 128–97°W), NE (34.4–52.1°N, 99–62°W), NW (34–52°N, 134–98°W) and SE (23–40°N, 99–68°W). Each has its own IC from the HRRR analysis and its own regridded GFS forcing, with LAND and OROG normalized by the CONUS statistics (Section 7). Winter case, 6-hour cadence to F48, scored against the native 3 km analysis; the 6 km CONUS run is scored over the same box against the decimated analysis.

Resolution invariance (all four nests). Each region is forecast three ways by the same 6 km HRRRCast checkpoint, all starting from the HRRR analysis: the native 3 km nest (the checkpoint run directly on the 530 × 900 window of the 3 km HRRR grid, with 3 km initial state, GFS forcing and statics, no decimation), a stand-alone 6 km run (the 6 km model run by itself on a 274 × 452-cell sub-domain cut from the CONUS 6 km inputs, not a cut-out of the CONUS forecast; the U-Net is fully convolutional and accepts any grid whose height + 2 and width are multiples of 4), and the CONUS run (the normal 530 × 900 CONUS forecast, cut to the same area afterwards). All three are scored on the same 6 km cells against the decimated analysis, so the 3 km / 6 km ratio isolates resolution and the stand-alone / CONUS ratio isolates the domain. No column is an HRRR physics-model forecast: all three are HRRRCast forecasts, and the HRRR analysis serves only as initial state and verifying truth.

HGT500 RMSE (m), F06 / F24 / F48	3 km nest (HRRRCast, 3 km input)	6 km stand-alone	CONUS run
SW	15 / 40 / 129	15 / 41 / 140	16 / 36 / 113
NE	267 / 594 / 663	267 / 596 / 663	19 / 58 / 214
NW	34 / 95 / 244	30 / 117 / 306	23 / 63 / 147
SE	98 / 277 / 337	109 / 229 / 281	15 / 40 / 116

MSLMA RMSE (hPa), F06 / F24 / F48	3 km nest (HRRRCast, 3 km input)	6 km stand-alone	CONUS run
SW	1.49 / 3.11 / 6.20	1.38 / 2.95 / 5.60	1.40 / 2.93 / 5.44
NE	7.12 / 25.05 / 30.13	6.90 / 23.86 / 28.73	1.18 / 4.72 / 15.95
NW	1.92 / 4.61 / 10.30	1.77 / 4.11 / 10.16	1.74 / 3.69 / 7.46
SE	7.24 / 21.22 / 50.50	8.30 / 29.89 / 51.41	1.04 / 2.94 / 7.47

Over F06–F48 and six scored fields (HGT500, MSLP, U10, HGT300, 900 hPa temperature and wind), the 3 km / 6 km error ratio stays within 0.91–1.19 (SW), 0.98–1.07 (NE), 0.79–1.15 (NW) and 0.68–1.22 (SE). Running the same areas stand-alone instead of inside CONUS changes the error by up to 2× (NW) and 3–23× (NE, SE). Resolution barely matters; the domain does. The NE nest fails identically at both resolutions. These results add to the evidence for near resolution invariance in our companion paper (in preparation). Our earlier native-3 km tests that suggested otherwise used the distorted statics of Section 7.

Caveats: one winter case and one ensemble member. For NW and SE, the 3 km GFS forcing was interpolated from the 6 km forcing rather than regridded natively; at SW the two differ by 0.013 standard deviations.

The GFS forcing is not the cause. The regridded GFS fields the model receives were verified against the HRRR analysis over each nest area (winter case; GFS winds rotated to the HRRR grid before scoring;

pressure-level cells below ground excluded). Their errors are small and similar everywhere, while the stand-alone nests' errors differ by up to 18× between regions:

GFS input RMSE, F06 / F48	HGT 1000 hPa (m)	HGT 850 hPa (m)	HGT 500 hPa (m)	HGT 250 hPa (m)	Wind 250 hPa (m/s)	T2M (K)	MSLP (hPa)
CONUS	5.8 / 6.1	5.2 / 8.3	5.7 / 11.5	7.9 / 17.5	3.1 / 7.2	2.3 / 2.2	1.6 / 1.4
SW	4.9 / 4.1	5.4 / 4.9	5.5 / 8.0	6.9 / 18.8	3.5 / 9.4	2.8 / 1.6	1.7 / 1.5
NE	5.9 / 7.5	4.6 / 12.1	6.2 / 16.9	7.3 / 23.3	2.8 / 7.4	2.2 / 2.8	1.7 / 1.4
NW	5.3 / 7.8	7.1 / 7.4	6.5 / 11.0	10.0 / 15.6	3.9 / 5.3	2.8 / 2.4	1.9 / 1.6
SE	6.2 / 5.3	3.5 / 4.9	4.2 / 10.1	7.3 / 12.7	2.3 / 7.0	1.6 / 1.9	0.8 / 0.7

At F06 the GFS 500 hPa height error is 4–7 m in every region, against stand-alone nest errors of 15 m (SW), 34 m (NW), 98 m (SE) and 267 m (NE). SE has the most accurate GFS input of all and the NE input is no worse than CONUS at F06; both nests still collapse. All four GFS levels (1000, 850, 500 and 250 hPa) and all 16 surface fields are present at every lead, with no missing or empty fields. The one deliberate substitution is precipitation: as in training, GFS precipitation is the 6-hour accumulation from the next synoptic-hour file, so it runs up to 3 hours ahead of the valid time. MSLP compares GFS PRMSL with HRRR MSLMA, whose sea-level reductions differ.

A' zones on the SW nest. With the A' design (widths in kilometres as at 6 km, capped at 150 cells = 450 km), F48 error relative to the nest without zones is HGT500 0.62, HGT300 0.64, HGT900 0.63, WIND300 0.69, TMP900 0.70, U10 0.80 and MSLP 0.86; the 50-cell edge bands reach 0.13–0.34 for HGT500. The interior beyond the zones is unchanged to F42 and 2–5% better at F48. MSLP is worse along the east, south and north edges at F18–F42 (up to 1.34), echoing the southern-edge MSLP weakness at 6 km. The nested A' forecast is within 0.84–1.19 of the 6 km A' run over the same area.

Where stand-alone nests fail. The NE nest collapses within the first 6-hour step (HGT500 error 267 m at F06 against 19 m for the CONUS run over the same area), although its IC and GFS forcing are identical to the CONUS inputs there. Running the 6 km model on 6 km sub-boxes of the CONUS inputs gives the same 270 m, so resolution plays no part. A series of one-step sub-box runs (winter, F06 HGT500 error relative to the CONUS run over the same cells) isolates two controls:

Sub-domain (rows × cols, 6 km)	West edge	West edge over ocean	Error ratio
274 × 452 SW	128–123°W	100%	0.94
274 × 452 NW	134–127°W	100%	1.30
274 × 452	124–119°W	9%	1.86
274 × 452	116–112°W	0%	5.30

Sub-domain (rows × cols, 6 km)	West edge	West edge over ocean	Error ratio
274 × 452	108–106°W	0%	9.84
274 × 452 NE	99°W	4%	13.86
274 × 644 (wider, to the Atlantic)	116–112°W	0%	3.09
402 × 452 (taller)	116–111°W	0%	3.17
402 × 644	116–111°W	0%	2.15
530 × 644 (full height)	116–110°W	13%	0.97

Small domains run well when their west edge is over the Pacific. With a land west edge, error grows the farther east the edge lies and shrinks as the domain grows, reaching the CONUS level only at full north–south extent. Widening toward the Atlantic helps no more than adding land to the south. The pattern is not atmospheric: the cause, identified below, is the network's global max pooling. The NE summer case fails the same way (ratio 8.7), so this is not a cold-domain effect. These are single-step results, mostly from one case.

Why stand-alone nests collapse: global max pooling. Physically, a 6-hour forecast for a 2700 km box given its true initial state should need little from outside: air travels about 1000 km at jet level and 150 km near the surface in 6 hours. The collapse is architectural. Each of the network's 47 CBAM channel-attention blocks (Woo et al. 2018) takes the global average and the global maximum of every feature over the whole domain and uses them to scale that feature at every grid point. Trained only on full CONUS, the network relies on CONUS-wide maxima; a sub-domain without the features that produce them receives wrong scalings everywhere at once. We tested this by intercepting the pooling layers for one 6-hour step of the NE 6 km box (same initial state, forcing and diffusion noise in every run) and substituting the values recorded at the same diffusion step of the full-CONUS run:

NE box, F06 RMSE	HGT500 (m)	MSLP (hPa)	U10 (m/s)	TMP900 (K)	HGT500 bias (m)
Full CONUS run, over the box	19.5	1.19	1.85	1.68	+7.6
Stand-alone box	268.1	6.85	14.50	14.28	−249
Box with all 94 global pools from the CONUS run	19.6	1.20	1.85	1.69	+7.8
Box with the 47 average pools from the CONUS run	275.1	6.78	15.47	15.46	−257

NE box, F06 RMSE	HGT500 (m)	MSLP (hPa)	U10 (m/s)	TMP900 (K)	HGT500 bias (m)
Box with the 47 max pools from the CONUS run	19.0	1.39	1.87	1.69	+0.9

Replacing the 47 global max pools alone restores the CONUS skill in every field; the average pools play no part. With the right maxima the box forecast equals the CONUS forecast over the same area, so the convolutions themselves are local enough for nesting. This explains every feature of the failure: it is immediate and domain-wide (NE HGT500 error is 242–272 m in every fifth of the box from west to east), identical at 3 km and 6 km (a maximum does not depend on grid spacing), and eased by larger or Pacific-bounded domains, whose maxima are closer to CONUS's. It is a shortcut learned from training on a single fixed domain, and the native 3 km model, which also uses CBAM, may share it.

Intuition. Each of the network's 47 attention blocks sets the volume of every learned feature from that feature's strongest value anywhere in the domain. A sub-domain that lacks the strongest features of CONUS weather, such as the Pacific jet in winter, gets wrong volumes everywhere at once, which a host supplying the CONUS maxima corrects. It is like a camera whose auto-exposure meters on the brightest spot: trained only on full-CONUS scenes, the network always had the “sun” in frame, and a crop that leaves it out is re-exposed wrongly as a whole. The gain is a sigmoid (0–1) per channel computed from the channel's global average and maximum by a small shared MLP, not a softmax over positions as in query–key attention. CBAM's second, spatial attention stage uses only a 7 × 7 neighborhood and is unaffected by the domain. The 47 blocks sit in residual blocks throughout the U-Net: 6 in the three input stems (HRRR state, GFS forcing, noised state), 10 in the two encoder levels, 20 at the bottleneck (1/4 resolution), 10 in the two decoder levels and 1 in the output refinement. These global pools are distinct from the U-Net's two local downsampling steps. The diffusion sampler is not the cause, since every forward pass of the U-Net computes these pools, but its 24 passes per 6-hour step let the wrong gains feedback on the features being pooled.

Why SW and NW run stand-alone. Recording the 47 max pools for one 6-hour step (F00 to F06, all 24 diffusion iterations) of the four 274 × 452 6 km boxes and of the CONUS run confirms that the Pacific-bounded boxes see maxima closest to CONUS's. The median relative distance ‖box − CONUS‖ / ‖CONUS‖ of the max-pooled channel vectors is 0.12 for SW and 0.14 for NW, against 0.18 for SE and 0.21 for NE. This is the same order as the F06 failure ratios (0.94, 1.30, 7.56, 13.86). In NE the distance grows from 0.19 at the first diffusion iteration to 0.28 at the last, as wrong gating feeds back on the features being pooled; in SW it stays near 0.13. The SW maxima are not identical to CONUS's, so the network tolerates differences of about 12–15% but not the 18–21% of the inland-bounded boxes. The average pools rank the boxes the same way (0.13, 0.17, 0.20, 0.22), so this measurement alone does not single out max pooling; the swap test above does.

Implication. Lateral relaxation cannot repair the collapse, but the network's own context channel can: running the nest in lockstep with a 6 km CONUS HRRRCast host and passing the host's global max pools to the nest is a HRRRCast-in-HRRRCast coupling that needs no retraining. Alternatively, training on random sub-domain crops would remove the dependence.

Host-coupled nests to 48 hours. The NE and SE nests, at 3 km (530 × 900 windows of the 3 km grid) and at 6 km (274 × 452 boxes), were run in lockstep with the 6 km CONUS HRRRCast host. At every 6-hour step the host takes its step first while its 47 global max pools are recorded at each diffusion step; each nest then takes the same step with those 47 layers returning the host's values. Everything else in a nest is its own: initial state, regridded GFS forcing, statics, diffusion noise, average pools and spatial attention. No lateral relaxation zones are used. The host reproduces the earlier CONUS run exactly.

RMSE, F06 / F24 / F48	NE HGT500 (m)	NE MSLP (hPa)	SE HGT500 (m)	SE MSLP (hPa)
CONUS run, over the area	19 / 58 / 214	1.18 / 4.72 / 15.95	15 / 40 / 116	1.04 / 2.94 / 7.47
3 km nest alone	267 / 594 / 663	7.12 / 25.05 / 30.13	98 / 277 / 337	7.24 / 21.22 / 50.50
3 km nest + host max pools	20 / 79 / 273	1.49 / 5.27 / 17.07	14 / 40 / 104	1.23 / 3.05 / 9.02
6 km box alone	267 / 596 / 663	6.90 / 23.86 / 28.73	109 / 229 / 281	8.30 / 29.89 / 51.41
6 km box + host max pools	19 / 82 / 269	1.38 / 5.17 / 15.67	14 / 39 / 108	1.09 / 2.63 / 7.05

Coupling removes the collapse at every lead and in every scored field (HGT500, MSLP, U10, HGT300, 900 hPa temperature and wind). SE stays at CONUS skill to 48 hours (F48 HGT500 0.90–0.93 of the CONUS run over the same area; MSLP 0.94 at 6 km, 1.21 at 3 km). NE tracks the CONUS run with a slow drift above it (F48 HGT500 1.26–1.28, MSLP 0.98–1.07); its error enters from the inland western edge, where the nest receives no boundary information, which A' zones remove (below). The 3 km / 6 km error ratio stays within 0.96–1.20 (NE) and 0.96–1.28 (SE) over all fields and leads, although the 3 km nests receive maxima computed by the 6 km host: the max-pooled values carry over across grid spacing, and near resolution invariance holds under coupling. One winter case, one member.

Coupled nests with edge zones. The same coupled NE and SE nests were rerun with lateral relaxation zones added at their four edges, applied to each nest's state after every 6-hour step. Two edge sources were compared. *GFS edges* are the A' zones of Sections 4–8: UGRD, VGRD, TMP, HGT and SPFH on all 20 levels plus the 10 m and 80 m winds, relaxed toward rotated GFS over the 6-hour advection widths, capped at 450 km (150 cells at 3 km, 75 at 6 km). *Host edges* relax all 136 model fields toward the CONUS HRRRCast host at the same hour, over the same widths. Host, max-pool coupling and everything else are unchanged from the run above, so the edges are the only difference. The GFS targets for the SE nest and the 6 km boxes were cut from the CONUS A' targets; on NE they match targets regridded directly from GFS to within 10^{-4} of the field's standard deviation.

RMSE, F06 / F24 / F48	NE HGT500 (m)	NE MSLP (hPa)	SE HGT500 (m)	SE MSLP (hPa)
CONUS run, over the area	19 / 58 / 214	1.18 / 4.72 / 15.95	15 / 40 / 116	1.04 / 2.94 / 7.47
3 km nest + host max pools	20 / 79 / 273	1.49 / 5.27 / 17.07	14 / 40 / 104	1.23 / 3.05 / 9.02
… + GFS edges	14 / 54 / 185	1.49 / 4.67 / 12.71	10 / 25 / 62	1.23 / 2.76 / 7.58
… + host edges	19 / 70 / 244	1.45 / 5.26 / 16.20	14 / 39 / 115	1.21 / 2.90 / 8.28
6 km box + host max pools	19 / 82 / 269	1.38 / 5.17 / 15.67	14 / 39 / 108	1.09 / 2.63 / 7.05
… + GFS edges	13 / 58 / 193	1.38 / 4.69 / 12.70	10 / 25 / 67	1.09 / 2.44 / 5.83
… + host edges	18 / 73 / 248	1.36 / 5.25 / 15.82	14 / 38 / 117	1.08 / 2.61 / 6.85

GFS edges remove NE's drift. At F48 the 3 km NE nest's HGT500 error falls from 1.28 to 0.86 of the CONUS run over the same area (−32%), and SE's from 0.90 to 0.54 (−40%). HGT300 is 0.81 (NE) and 0.55 (SE), U10 0.84 and 0.98, and MSLP 0.80 and 1.01. The gain appears from F06 and holds at every lead in every scored field. The 6 km boxes behave the same, and the 3 km / 6 km ratio stays within 0.94–0.96 for F48 HGT500, so near resolution invariance holds with zones as well. The weakest fields remain at low levels, where the zones are narrowest: SE 900 hPa wind (1.21 of CONUS at F48) and NE 900 hPa temperature (1.11). Host edges give only about a 10% gain (F48 HGT500 1.14 NE, 0.99 SE), as expected: relaxing toward the CONUS HRRRCast cannot do better than that forecast, whereas GFS brings independent information across the boundary, the role lateral boundary conditions play for a physics nest. Combined, the two retraining-free methods — host max pools for global context and height-dependent GFS zones for boundary information — give 3 km nests that beat the CONUS run over their own area at 48 hours (HGT500 0.54–0.86). One winter case, one member.

One-way nesting with an HRRRCast host. The SW nest was relaxed toward the 6 km HRRRCast A' host forecast instead of GFS, with the same network for host and nest, as physics nested models use the same parameterizations in every domain. Because host and nest share the HRRR terrain, all 136 model fields are relaxed, including surface pressure, 2 m temperature and MSLP. The host zones improve the nest (F48 HGT500 0.75 of the nest without zones) but trail GFS zones (1.22 × the A' nest). The difference is by edge: the nest's west and south edges coincide with the CONUS boundary, where the host is itself relaxed to GFS, and there the two nests match; at its east and north edges, inside CONUS, the host's own error enters (east-edge HGT500 6.6 × the A' nest at F48). A host-relaxed NE nest is not rescued (F06 HGT500 still 13.8 × the host over the same area), as expected for a failure that occurs in the first step.

HRRRCast-in-HRRRCast. Replacing the nest's 40 GFS input channels with fields built from the host forecast (250 hPa interpolated, winds rotated back to earth-relative, 12 km smoothing, GFS normalization) raises the SW nest's error by 13–47% from F06 onward (HGT500 1.28 and MSLP 1.32 at F06, relative to the host-zone nest with GFS inputs). One-step ablations place the penalty in the mass and wind channels; vertical-velocity, moisture and cloud inputs are neutral. With mass and wind inputs

taken from the verifying HRRR analysis, a perfect host, the penalty largely disappears but the nest only ties GFS (MSLP 0.98, U10 0.96, HGT500 1.09). With this checkpoint a HRRRCast host can at best equal GFS-driven nesting, and only where it is as accurate as GFS; exploiting a better host requires training with host-built inputs.

11. Discussion and conclusions

A simple, retraining-free boundary treatment fixes HRRRCast's main long-lead weakness, but it works differently from physics-model boundary conditions.

1. **Size zones to the model's cadence, not a physics time step.** In a model with a 6-hour cadence the zone must span the 6-hour advection distance, which makes upper-level zones wide (97–180 cells).
2. **Relax the full column and the near-surface winds.** Partial-level relaxation and omitting surface winds both degrade the fields left alone.
3. **Relax all around the boundary.** Unlike a physics model, the network's edge errors are not inflow-driven, and outflow relaxation helps too.
4. **Check coordinate frames between global and regional inputs.** A fixed rotation between GFS and HRRR winds survived into an operational AI model because the network learned around it.
5. **Zones alone can replace the full-domain GFS inputs.** With the four-level GFS inputs zeroed and GFS entering only through the A′ zones, the winter F48 forecast beats current HRRRCast (full GFS, no zones) for heights and temperatures (HGT500 0.80, HGT300 0.71, TMP300 0.72 of baseline). It matches it for MSLP (1.01) and is worse for winds (U10 1.15). It trails A′ by up to 0.2 in these ratios at F48 and is 1.2–3.3× baseline through F24, the cost of a checkpoint never trained this way. A nested model without interior GFS can therefore run on zones alone; training with boundary-only forcing should remove most of the early penalty (one winter case so far).
6. **Normalize static fields with fixed training statistics.** Statistics computed from the input domain silently change the terrain the network sees on any sub-domain; on a nested domain this alone can ruin the forecast (Section 7).
7. **Couple small nests to a CONUS host, and relax their edges toward GFS.** Stand-alone sub-domains whose western edge lies over land collapse because the network's global max pools expect CONUS-wide maxima. Replaying the host's max pools removes the collapse without retraining, and adding A′ zones at the nest edges brings the 3 km nests below the CONUS run's own error over the same area at 48 hours (Section 10).

Relation to the new native 3 km HRRRCast. GSL's latest model (Abdi et al. 2026b) is trained at native 3 km on nine years of data, using the same building blocks as the 6 km checkpoint studied here (CBAM attention and separate input branches); four points carry over:

- **Same coupling design.** It is trained on 1-, 3- and 6-hour leads, but the public rollout (`rollout_hour = 6` in `src/fcst.py`, master `b81da9a`) still advances the state at a 6-hour cadence of direct steps, with GFS conditioning at 250, 500, 850 and 1000 hPa over the whole domain. Zone widths should therefore still span 6-hour advection, which on the 3 km grid is twice as many cells (Section 10 capped them at 150 cells).
- **Both preprocessing issues remain in the released code.** `make_bcs.py` still regrids GFS winds without rotating them (Section 7), and the 3 km model's `normalize-stats.nc` (48 variables)

still has no LAND or OROG statistics, so any sub-domain inherits the static-field distortion of Section 7.

- **Resolution invariance is a large-scale result.** Abdi et al. show that native 3 km training improves fine-scale skill (fractions skill score, spectra) because 6 km subsampling aliases the smallest scales. Our finding is complementary: run on native 3 km input, the 6 km model keeps its large-scale skill (HGT500, MSLP, winds), but it cannot supply fine-scale structure it never learned.
- **The domain effect should be tested on the 3 km model.** It too is trained only on full CONUS, so stand-alone sub-domains with an inland west edge may fail in the same way. The four-region test of Section 10 is the natural first check, alongside A' zones at 3 km.

Smith et al. (2026) describe a nested AI model that matches the full vertical structure of its driving model; HRRRCast's four-level GFS conditioning and the full-column zones studied here sit between that approach and the conventional physics-model relaxation zone.

Limitations. Two cases, one ensemble member each, and the nesting experiments use only the winter case; no physics-model comparison beyond F18; upper-level zones cover much of the domain, so part of the upper-air gain is GFS itself; the model's full-domain GFS conditioning remains (Section 9).

Related work. Relaxation boundary conditions follow Davies (1976); Warner et al. (1997) review lateral boundary conditions as a limitation of regional NWP. AI limited-area models include HRRRCast (Abdi et al. 2026a, 2026b), Neural-LAM (Oskarsson et al. 2023) and a surface-variable AI limited-area model in *Communications Earth & Environment*. The wind-frame mismatch and the max-pooling dependence are forms of shortcut learning (Geirhos et al. 2020) and Clever Hans predictors (Lapuschkin et al. 2019); climate-invariant machine learning (Beucler et al. 2024) addresses input representations that generalize beyond the training distribution. NWP practice for rotating winds between grids is documented for wgrib2 `-new_grid_winds` and for rotating WRF winds.

Data and code. Checkpoint `3355614` and preprocessing from NOAA-GSL/HRRRCast-live; GFS 0.25° and HRRR analyses from the NOAA Open Data Dissemination buckets on AWS (`noaa-gfs-bdp-pds`, `noaa-hrrr-bdp-pds`). Experiment scripts: `fast_forecast_relax.py`, maxpool_nest.py and maxpool_relax_nest.py, with their scoring tools (to be released).

Acknowledgments. We wish to recognize Professor Karuna Joshi and the support provided by the UMBC NSF-funded Center for Accelerated Real Time Analytics (CARTA) for providing their computing and laboratory resources to support this research study. We want to express our gratitude to the NASA/ESTO FireTech Program Managers, Teresa Kauffman, Haris Riris (Ret.) for their support to our Grant 80NSSC22K1405. We wish to thank NSF for NAIRR access to PSC computing resources. Finally, we thank the UMBC OIT for access to the CHIP computing resources.

References

Abdi, D., I. Jankov, P. Madden, V. Vargas, T. A. Smith, S. Frolov, M. Flora, and C. Potvin, 2026a: HRRRCast: A data-driven emulator for regional weather forecasting at convection allowing scales. *Artificial Intelligence for the Earth Systems*, **5**(2), 250061, https://doi.org/10.1175/AIES-D-25-0061.1 (arXiv:2507.05658).

Abdi, D., I. Jankov, V. Vargas, P. Madden, E. Carpenter, T. Wilson, J. Hamilton, M. Smith, L. Blank, L. Cui, J. Wang, and M. Mahalik, 2026b: Towards operational data-driven regional weather forecasting: Extreme-event skill and near-real-time verification with HRRRCast. arXiv:2610.36186.

Beucler, T., and Coauthors, 2024: Climate-invariant machine learning. *Science Advances*, **10**, eadj7250, https://doi.org/10.1126/sciadv.adj7250.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533–538.

Black, T. L., and Coauthors, 2021: A limited area modeling capability for the Finite-Volume Cubed-Sphere (FV3) dynamical core and comparison with a global two-way nest. *Journal of Advances in Modeling Earth Systems*, **13**, e2021MS002483, https://doi.org/10.1029/2021MS002483.

Bodnar, C., and Coauthors, 2025: A foundation model for the Earth system. *Nature*, **641**.

Davies, H. C., 1976: A lateral boundary formulation for multi-level prediction models. *Quarterly Journal of the Royal Meteorological Society*, **102**, 405–418.

Dowell, D. C., and Coauthors, 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Weather and Forecasting*, **37**, 1371–1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Geirhos, R., J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, 2020: Shortcut learning in deep neural networks. *Nature Machine Intelligence*, **2**, 665–673.

Hamer, S., J Sleeman, M.Halem, Towards a Dynamic Data Driven AI Regional Weather Forecast Model. Springer Cham,2025,Pages XVI, 443, 126-134,

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416–1421.

Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, 2019: Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, **10**, 1096.

Oskarsson, J., T. Landelius, and F. Lindsten, 2023: Graph-based neural weather prediction for limited area modeling. *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*, arXiv:2309.17370.

Pathak, J., and Coauthors, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv:2202.11214.

Schmude, J., and Coauthors, 2024: Prithvi WxC: Foundation model for weather and climate. arXiv:2409.13598.

Skamarock, W. C., and Coauthors, 2019: A description of the Advanced Research WRF Model Version 4. NCAR Technical Note NCAR/TN-556+STR, National Center for Atmospheric Research.

Smith, T. A., M. Pope, S. Frolov, B. Basarab, D. Abdi, P. Madden, and I. Jankov, 2026: Bridging short- and medium-range weather forecasting with machine learning. arXiv:2608.26822.

Song, J., C. Meng, and S. Ermon, 2021: Denoising diffusion implicit models. *International Conference on Learning Representations (ICLR)*, arXiv:2010.02502.

Warner, T. T., R. A. Peterson, and R. E. Treadon, 1997: A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bulletin of the American Meteorological Society*, **78**, 2599–2617.

Woo, S., J. Park, J.-Y. Lee, and I. S. Kweon, 2018: CBAM: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19.