Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

WeatherBridge and the limits of global weather interpolation at held-out query times

Daniil Sukhorukov

suhorukov.dan0@gmail.com

AIRI - Artificial Intelligence Research Institute

Andrei Zakharov

AIRI - Artificial Intelligence Research Institute

Yuri Maximov

International Center for Corporate Data Analysis

Ilya Makarov

AIRI - Artificial Intelligence Research Institute

Ivan Oseledets

AIRI - Artificial Intelligence Research Institute

Article

Keywords: weather interpolation, held-out query times, IFS HRES, anchor transport, ERA5

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11183024/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

WeatherBridge and the limits of global weather interpolation at held-out query times

Daniil Sukhorukov[1,2*], Andrei Zakharov[1], Yuri Maximov[3], Ilya Makarov[1,4], Ivan Oseledets[1]

[1]AI Research Institute, Moscow, Russia.
[2]AI Talent Hub, ITMO University, Saint Petersburg, Russia.
[3]International Center for Corporate Data Analysis, Astana, Kazakhstan.
[4]Trusted AI Center, RAS, Moscow, Russia.

*Corresponding author(s). E-mail(s): suhorukov.dan0@gmail.com;

Abstract

Many forecast archives provide six-hourly fields, but applications need hourly states. We ask whether a learned interpolator can recover intermediate trajectories from two endpoints at query hours omitted from training. We compare linear interpolation and six architectures at six- and twelve-hour horizons, and test transfer to IFS HRES forecasts. A matched six-hour WeatherBridge control receives supervision at every interior hour. WeatherBridge combines query-conditioned bidirectional transport with multiscale and spectral corrections. It has the lowest aggregate error at held-out query hours at both horizons, although no model wins every field. The matched control improves held-out query hours but barely changes trained ones, measuring the cost of withholding query times. Models trained on ERA5 lose this advantage on forecast trajectories; HRES adaptation recovers about 7% over linear interpolation in a retrospective audit. Learned interpolation captures nonlinear motion and small-scale structure between sparse anchors, while temporal coverage and anchor-domain mismatch bound its benefit.

Keywords: weather interpolation, held-out query times, IFS HRES, anchor transport, ERA5

1 Introduction

Global forecast archives and benchmarks commonly expose states on fixed temporal grids. GraphCast predicts at six-hour intervals [1], whereas Pangu-Weather includes separate one-, three-, six- and 24-hour models [2]. FuXi, Aurora, Stormer, AIFS and GenCast provide further examples of learned global forecast systems [3–7]. FourCastNet uses Fourier operators for global prediction [8], while NeuralGCM learns components of a numerical general circulation model [9]. ClimaX trains a foundation model spanning weather and climate tasks [10]. FengWu extends skilful deterministic forecasts beyond a ten-day lead for selected fields [11], while Aardvark predicts weather directly from observations [12]. Keisler demonstrated an early global graph-network forecaster [13]. WeatherBench and WeatherBench 2 provide shared forecast-verification protocols [14, 15]. ERA5 provides hourly reference states [16]. Recovering an interior state x_τ from known anchors x_0 and x_T is an interpolation problem, not a forecast: information from the later anchor is available by definition. Linear interpolation is consequently a strong reference, but it assumes that each Eulerian grid value follows the chord between its endpoint values. A translated pressure system, moisture filament or near-surface wind feature does not obey that assumption. Its amplitude can be attenuated even when the two anchors resolve the feature clearly. Motion-aware reconstruction has precedents in optical-flow precipitation nowcasting [17, 18] and, more broadly, in the characteristic-based transport used by semi-Lagrangian atmospheric schemes [19].

Figure 1 shows the interpolation task and the target-dependent training loop. The interior ERA5 state provides supervision during training and the reference during evaluation; inference uses only the two anchors and query metadata. We use 2020 for development, 2021 for calendar transfer and a prespecified 2022 ERA5 sample for a later calendar check. Methods records the complete split and HRES development chronology.

Weather time interpolation has received less controlled study than weather forecasting. ModAFNO [20] adds τ conditioning to an Adaptive Fourier Neural Operator;

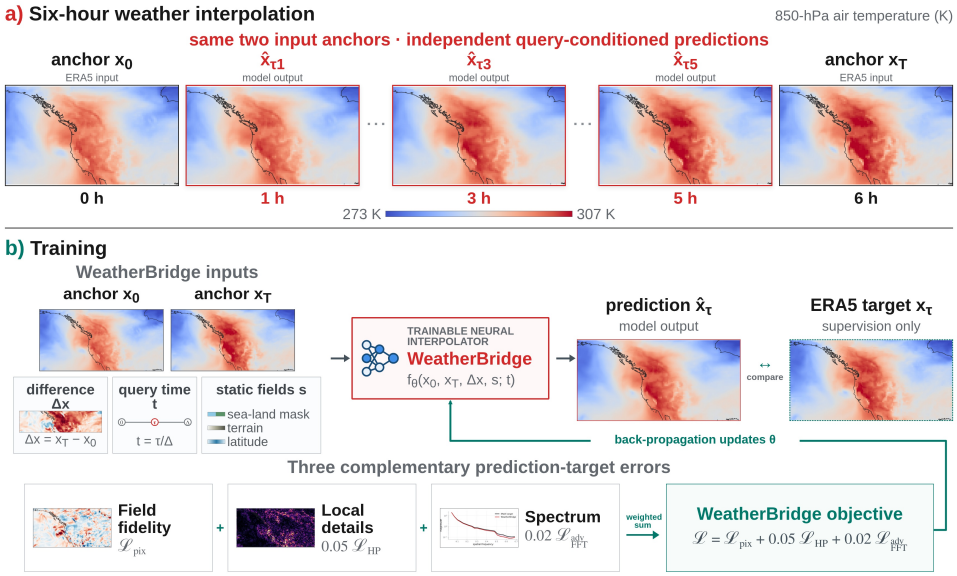


Fig. 1 WeatherBridge interpolation and training workflow. **a)** Six-hour reconstruction of 850-hPa air temperature during the Pacific Northwest heatwave of 28–29 June 2021. The same anchors condition independent predictions at each query time; ellipses denote the omitted second and fourth hours. **b)** Training uses the anchors, their difference, static fields and $t = \tau/\Delta$. The interior ERA5 state supplies only the pixel, high-pass and spectral error terms. The green path is active during back-propagation and absent at inference.

DYffusion [21] couples interpolation and forecasting in a probabilistic sampler. Its diffusion process uses temporal interpolation in place of the Gaussian noising process of denoising diffusion models [22]. Spherical DYffusion applies this framework to global climate emulation with spherical operators [23]. SpateGAN-ERA5 [24] instead addresses generative spatial and temporal downscaling of precipitation. Recent task-specific systems extend the range: HFD-CF [25] combines variable-specific flow with multiresolution warping for four midpoint fields, MSF-TCMA [26] combines multi-scale wavelet and multi-source features with a continuous-time manifold on regional unseen-time tests, while HourGlass [27] reconstructs probabilistic, temporally coherent trajectories from forecast-system output. Their variables, resolutions, training sources and query-time protocols differ, so published scores do not resolve the architectural comparison or behaviour at longer anchor intervals. Generative nowcasting models

learn future precipitation distributions [28, 29], neural weather models such as MetNet forecast precipitation directly from observations [30], and CorrDiff reconstructs fine spatial structure conditional on coarse atmospheric fields [31].

Video-frame interpolation suggests three useful mechanism families. Direct synthesis avoids an explicit flow field [32]; VFIformer and EMA-VFI aggregate long-range or inter-frame evidence with attention [33, 34]; and DAIN, RIFE, IFRNet and AMT estimate motion before fusing transported endpoints [35–38]. Quadratic Video Interpolation and EQVI add acceleration for non-uniform motion [39, 40], while softmax splatting treats collisions during forward warping [41]. Super SloMo makes intermediate time an explicit input [42]; FILM handles large image motion with multiscale features [43]. Diffusion models have also been applied to frame interpolation [44, 45].

Global atmospheric fields change the requirements behind these mechanisms. Longitude is periodic, grid cells have unequal area, the poles are coordinate singularities, and 24 variables must remain mutually consistent. Spectral power can look correct even when a front is displaced, whereas a smoother field can reduce pointwise error while erasing an informative residual. We therefore evaluate transport together with spherical scalar and vector spectra, diagnostic balances and temporal consistency. WeatherBridge adapts second-order transport to this setting; the quadratic displacement formula itself follows prior VFI work.

We compare explicit anchor transport with additive encoder–decoder reconstruction under a common 24-field protocol. The benchmark includes a time-conditioned SwinV2 [46], ModAFNO, S-DYff with Spherical FNO [47], PixelAttn-VFI, the DC-AE-derived WeatherDCAE [48], and WeatherBridge. SwinV2 and WeatherDCAE represent direct residual synthesis, PixelAttn-VFI represents feature-level endpoint interaction, and WeatherBridge supplies explicit transport. These are weather-specific implementations inspired by the cited model families rather than reproductions of image-domain systems.

Because their objectives and update budgets differ, the benchmark measures paper-scale performance rather than matched-compute efficiency.

Supplementary Fig. S14 shows how WeatherBridge combines anchor transport with additive reconstruction. Transport moves existing weather structures towards the query state; additive branches represent changes in their amplitude and shape. A learned gate combines transported anchors with linear interpolation, and a final temperature-column head adjusts geopotential and mean sea-level pressure. The study makes three contributions, each tied to a separate experiment:

1. **Held-out query-hour benchmark.** Six weather-adapted architectures share a 24-field protocol. A matched WeatherBridge control trained on every interior hour isolates query withholding from intrinsic hour difficulty. This supports transfer to sampled unseen queries, not continuous-time generalisation; the 12 h experiment remains an interval-length ablation.
2. **WeatherBridge.** A linear residual scaffold is combined with bidirectional second-order transport, multiscale corrections and spectral supervision. The model gives the lowest aggregate error at held-out query hours at both anchor intervals, with field-level exceptions; same-checkpoint lesions test acceleration, warped-anchor transport and the hydrostatic head.
3. **Forecast-trajectory transfer.** ERA5-trained models lose their advantage on IFS HRES anchors. Limited HRES adaptation restores an aggregate gain over linear interpolation, separating interpolation skill from anchor-domain mismatch; native data-driven forecast trajectories remain untested.

2 Results

We evaluate aggregate accuracy together with fieldwise, spectral, event-scale and physical diagnostics. ERA5 reconstruction and interpolation between IFS HRES forecast states answer different questions and are reported separately.

2.1 Interpolation at held-out query hours

Fig. 2 shows normalised root-mean-square error (RMSE) and anomaly correlation coefficient (ACC) versus τ on the 2020 development year. WeatherBridge is the proposed transport model, and WeatherDCAE-14M is the compact deep-compression reference under the same 24-channel scoring protocol. Error is largest near the midpoint, while the two held-out hours retain the same ordering as their trained neighbours. Table 1 separates trained from held-out query coordinates at both anchor gaps. The aggregate RMSE is 0.06263 for WeatherBridge, 0.07075 for WeatherDCAE-14M and 0.12680 for Linear Interp. The per-field results below resolve these averages into gains and regressions.

At the held-out hours of the six-hour task, WeatherBridge reaches RMSE 0.07431, 9.0% below WeatherDCAE-14M and 48.6% below Linear Interp. At the held-out hours of the twelve-hour task, its RMSE is 0.13277, improvements of 5.8% and 52.3% over

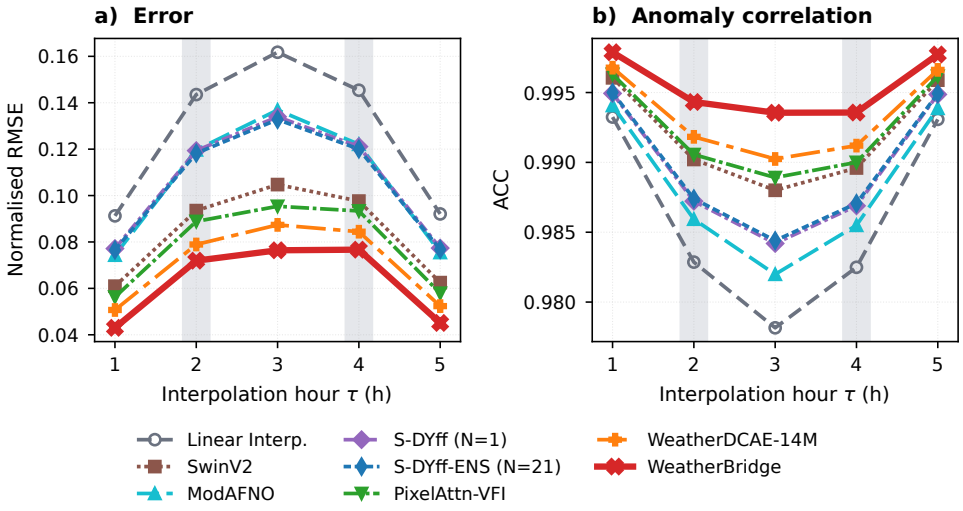


Fig. 2 Full-year 2020 scores for the principal models. **a)** Normalised RMSE. **b)** ACC against the WeatherBench-2 1990–2019 hourly climatology. WeatherBridge is the emphasized red curve. Grey bands mark the held-out query hours $\tau \in \{2, 4\}$, which are absent from training. Curves share a window index; tests use paired seven-day blocks. S-DYff and S-DYff-ENS use one and 21 stochastic reconstructions, respectively, from the same checkpoint.

Table 1 Performance at trained and held-out query hours on the 2020 ERA5 benchmark. Six-hour models train at $\tau \in \{1, 3, 5\}$ and hold out $\{2, 4\}$; twelve-hour models hold out $\{4, 6, 8\}$. Entries are mean ± standard deviation of the 24-channel macro RMSE across query hours. The standard deviation describes variation between hours, not sampling uncertainty over dates. Differences between the trained and held-out columns also reflect the intrinsic difficulty of their query hours and are not an estimate of a generalisation gap.

Model	6 h seen	6 h held out	12 h seen	12 h held out
WeatherBridge	$\mathbf{0.055 \pm 0.019}$	$\mathbf{0.074 \pm 0.003}$	$\mathbf{0.099 \pm 0.033}$	$\mathbf{0.133 \pm 0.008}$
WeatherDCAE-14M	$\underline{0.063 \pm 0.021}$	$\underline{0.082 \pm 0.004}$	$\underline{0.107 \pm 0.034}$	$\underline{0.141 \pm 0.008}$
Linear Interp.	0.115 ± 0.040	0.144 ± 0.001	0.202 ± 0.074	0.278 ± 0.015

the same references. The gain over Linear remains close to the gain at trained hours, and the model ordering persists at the held-out query hours.

A matched six-hour self-comparison measures the cost of withholding hours 2 and 4. Direct supervision lowers WeatherBridge RMSE at those hours by 8.77% in 2020 and 8.78% in 2021; at trained hours 1, 3 and 5, the change is below 0.4% (Supplementary Fig. S8). Supplementary Fig. S9 resolves the change by field and hour, and Supplementary Table S4 records which checkpoint comparisons match both seed and update budget. The sparse-query checkpoint transfers useful skill to both omitted hours; direct supervision retains an 8.8% advantage. This test covers integer hours, and the twelve-hour setting lacks an all-query control.

Figs. 3, 4 and 5 resolve the six-hour aggregate into the five interior-hour trajectories for every evaluated field. WeatherBridge has the lowest mean RMSE over $\tau = 1, \ldots, 5$ for 19 of 24 fields. WeatherDCAE-14M remains better for the four geopotential levels and MSLP. WeatherBridge wins most fields, but the mass-field regressions rule out a field-uniform claim.

The S-DYff-ENS curves add ensemble averaging to the same checkpoint and evaluation windows. Supplementary Figs. S15, S16 and S17 complete the field-wise comparison for twelve-hour interpolation in 2021 and both intervals on the fixed 2022 sample. The ensemble-size comparison and its paired single-draw control are described in Supplementary Section 17.

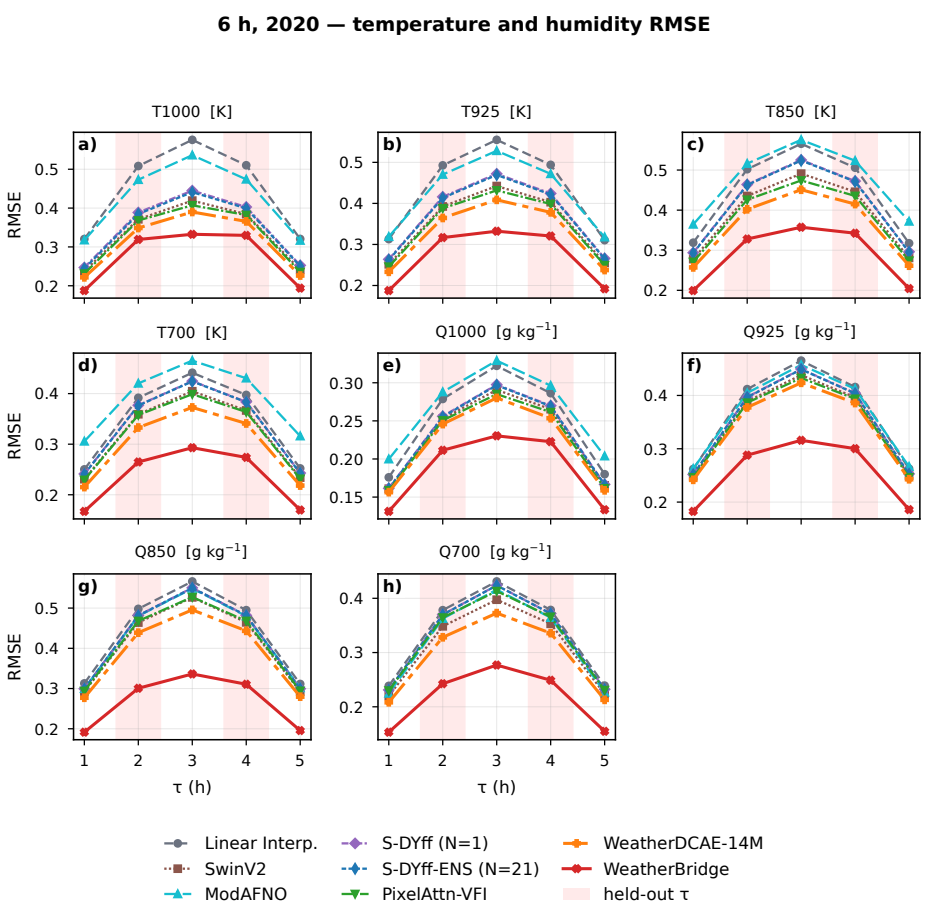


Fig. 3 Full-year 2020 physical-unit RMSE trajectories for temperature and specific humidity at four pressure levels. Each panel follows the six-hour interpolation from $\tau = 1$ to 5 h; lower is better. Pink bands mark held-out query hours $\tau \in \{2, 4\}$. WeatherBridge is the solid red curve. Scales differ between fields and should only be compared within a panel. All neural curves use the same evaluator and full-year window index.

Supplementary Figs. S3–S4 report the corresponding 12-hour ablation without repeating the six-hour panels above. Units and climatological scales differ, so comparisons are meaningful within a field and query hour, not across variables. Fig. 2a gives the normalised aggregate; Supplementary Figs. S6–S7 show the 2021 field ordering.

On the 2022 sample, WeatherBridge has lower pointwise RMSE in 109 of 120 six-hour and 254 of 264 twelve-hour field–hour cells. The paired aggregate and held-out intervals in Supplementary Table S5 exclude zero at both gaps. Held-out RMSE is

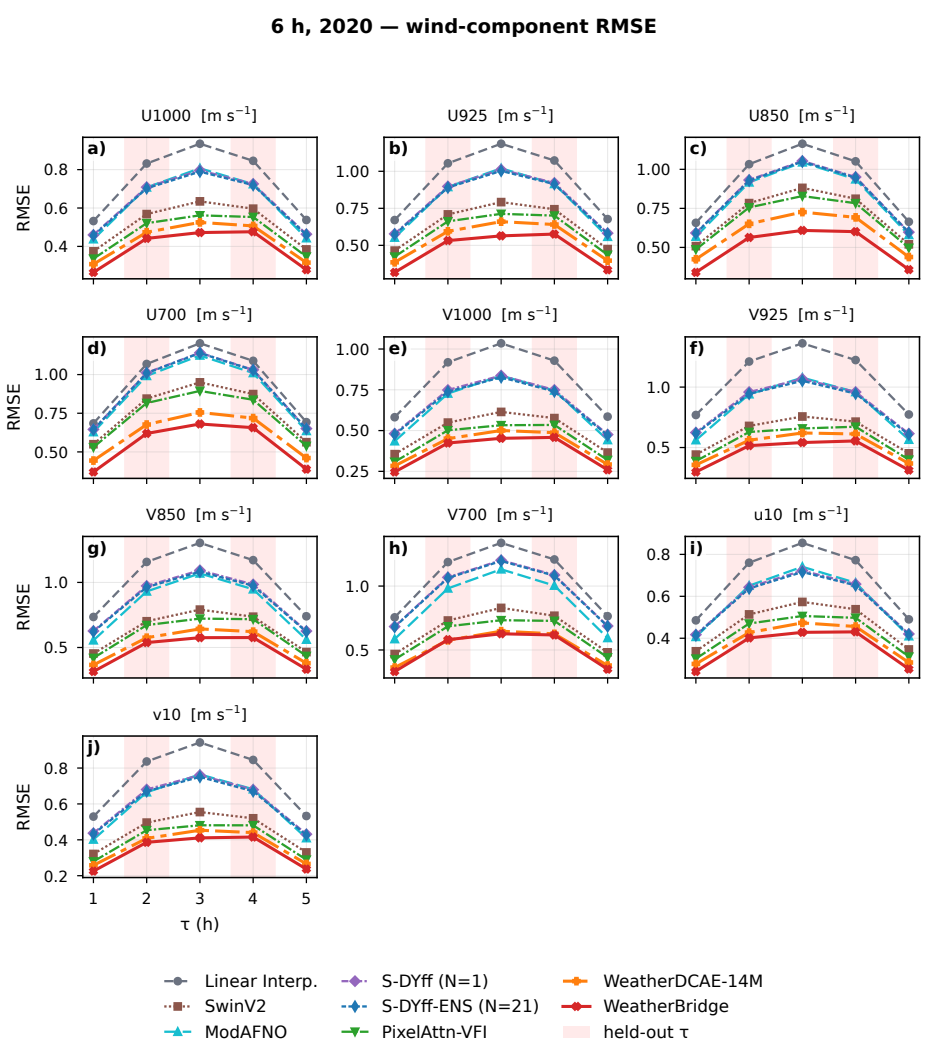


Fig. 4 Full-year 2020 physical-unit RMSE trajectories for zonal and meridional wind at four pressure levels and 10 m. Plotting conventions and model provenance match Fig. 3.

11.09% lower than WeatherDCAE-14M at six hours and 5.69% lower at twelve hours. The internal Detail and Refine variants are 0.3–0.4% better at 6 h in 2020 and 2021, but WeatherBridge is better at 12 h and has the lowest aggregate score when the two gaps are considered together. Supplementary Table S9 reports that comparison.

The sample preserves the aggregate and held-out ordering while revealing Holm-significant regressions in ten six-hour and seven twelve-hour field–hour cells. The

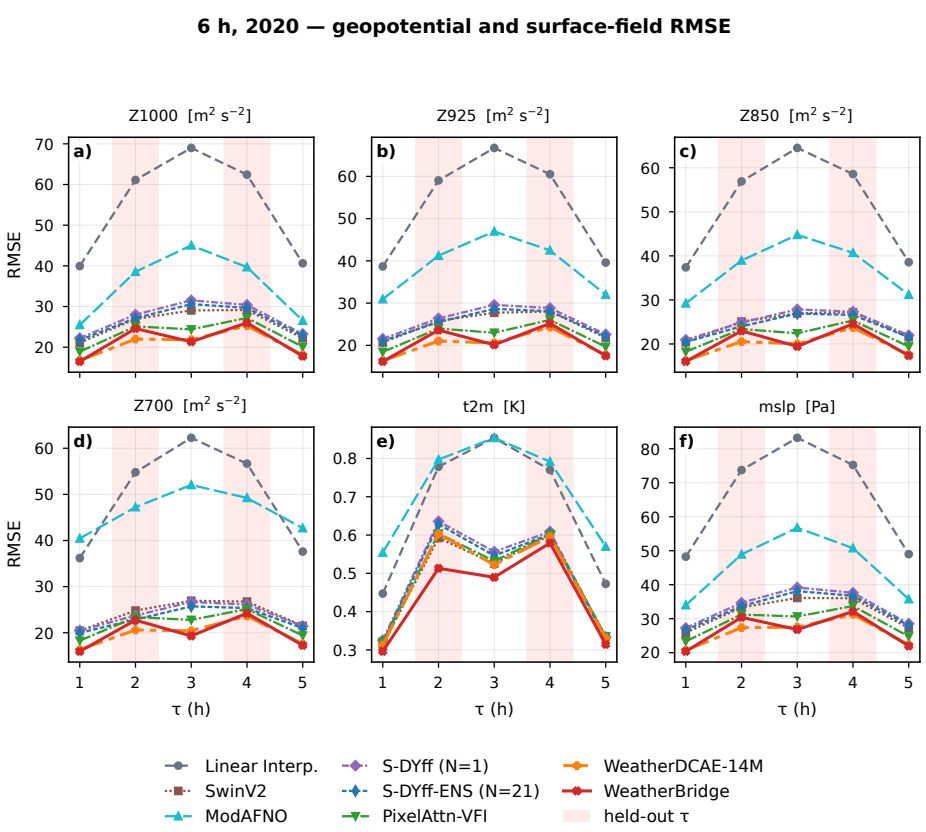


Fig. 5 Full-year 2020 physical-unit RMSE trajectories for geopotential at four pressure levels, 2-m temperature and mean sea-level pressure. Plotting conventions and model provenance match Fig. 3.

internal robustness sensitivity consequently retains WeatherDCAE-14M; we report that sensitivity analysis in Supplementary Table S7 rather than using it to define the study's central held-out endpoint. Supplementary Table S8 gives representative vector-SHT comparisons.

Same-checkpoint lesions show that both transport terms contribute to the reported checkpoint. Disabling acceleration raises aggregate RMSE by 2.16%; disabling warped-anchor transport raises it by 60.54% on the full-year 2020 index. Disabling the hydrostatic head has a smaller aggregate effect (+0.04%), but worsens all 25 targeted geopotential/MSLP field–hour cells after Holm correction; mean hydrostatic-balance MSE rises by 1.06% across the five query hours (Supplementary Table S10). These

lesions test whether the trained checkpoint uses each term, not how a model trained without that term would perform. The hydrostatic head has a small field-specific effect, whereas transport accounts for the largest checkpoint sensitivity. The fine, coarse and SFNO residual branches and the learned blend gates are not isolated by these tests, so their individual accuracy gains remain unquantified. The retrained Detail and Refine variants change several mechanisms and do not resolve that attribution. The 2021 comparison reuses the model set, normalisation, full-year window construction and query-hour split from 2020; no checkpoint trains on 2021. Because it informed development, the 2021 result is a calendar-transfer diagnostic. The prespecified 2022 sample supplies the later calendar check. Supplementary Table S3 reports its aggregate six-hour scores.

Exchanging the two anchors and replacing t by $1 - t$ produces a different prediction. On the retained 48-endpoint diagnostic, WeatherBridge's exchange RMSE is 0.287 in 2020 and 0.285 in 2021, more than three times its forward reconstruction error (Supplementary Table S11). The transport heads are bidirectional, while the complete interpolator is asymmetric because the left and right endpoints retain different learned roles.

WeatherBridge improves on WeatherDCAE-14M at every UTC start in both years. Its pooled advantage is 12.7% within one assimilation window and 9.7% across a boundary in 2020, with corresponding 2021 values of 12.6% and 9.6% (Supplementary Table S6). Assimilation timing changes the size of the gain, which nevertheless persists in all eight year–start strata.

2.2 Transfer to IFS HRES forecast trajectories

We replace the ERA5 anchors with two states from one IFS HRES forecast trajectory and retain ERA5 as the intermediate target. For each anchor pair, $\tau = 1, \ldots, 5$ denotes the offset inside the six-hour interval; forecast lead is a separate coordinate. The

resulting error combines temporal interpolation with HRES–ERA5 mismatch. The twelve-hour same-trajectory scores by field group and illustrative lead-bin trajectories are in Supplementary Table S2 and Supplementary Fig. S5, respectively.

Full-weight adaptation across forecast leads

Unlike the coefficient adapter below, this experiment updates the network weights. Its 24-initialisation 2022 index is distinct from both the 16-initialisation coefficient-adapter test and the 48-window ERA5 calendar-transfer sample. The row labelled "fine-tuned WeatherDCAE-14M" in Table 2 denotes full-weight optimisation, not the WeatherBridge coefficient adapter.

Lead coverage was inspected on the 2021 development archive. Fine-tuning on the short deployment range improves mean field–hour RMSE over linear interpolation by 1.70%. Extending training to 0–216 h on a 24 h lead grid raises the gain to 7.63%; using the native 6 h lead grid raises it to 12.35%.

The two full-range checkpoints specialise by lead. Routing the coarse-grid checkpoint below 48 h and the dense-grid checkpoint thereafter gives mean RMSE 0.4671, compared with 0.5567 for Linear, 0.5143 for the coarse checkpoint and 0.4880 for the dense checkpoint. Relative to each comparator, the routed result improves by 16.10%, 9.18% and 4.29%, respectively. These development results define the routing rule but provide no confirmation evidence.

On the fixed 2022 dates, the full-range coarse-grid checkpoint reduces the mean of the 120 field–hour RMSE values from 0.2902 to 0.2691 (7.28%). The lead-routed ensemble reaches 0.2675 (7.83%), but its spectral promotion audit is incomplete. We therefore retain the coarse-grid checkpoint as the adapted model of record and report the ensemble as a secondary result. Because the full-range branch postdates inspection of 2022, these values are retrospective rather than independent confirmation evidence.

Statistical resampling uses the 24 forecast initialisations as blocks; the 600 window–hour records remain grouped within them.

Table 2 Six-hour interpolation from IFS HRES forecast anchors to ERA5 on a fixed 2022 retrospective set (24 initialisations; left-anchor leads $\{0, 24, 48, 72, 96\}$ h; right endpoint through 102 h). Scores are the arithmetic mean of 120 latitude-weighted field–hour RMSE values. This index is distinct from the coefficient-adapter and ERA5 calendar-transfer samples. The lead-routed ensemble gives the lowest point estimate, but its spectral promotion audit has not been run; bold therefore marks the coarse-stride model of record, which passes the separately reported spectral criterion. The full-range branch postdates inspection of 2022, so this table is descriptive rather than independent confirmation.

Method	RMSE	vs. linear
Linear interpolation	0.2902	—
WeatherBridge, no adaptation	0.2866	+1.23%
WeatherDCAE-14M, no adaptation	0.2876	+0.91%
WeatherDCAE-14M, short-range fine-tuned	0.2740	+5.60%
WeatherBridge, short-range fine-tuned	0.2711	+6.59%
WeatherBridge, full-range coarse grid	**0.2691**	**+7.28%**
WeatherBridge, full-range dense grid	0.2715	+6.45%
WeatherBridge, lead-routed ensemble	0.2675	+7.83%

The corresponding spectral audit covers Q850 and U850 at query hours 2 and 3 over spherical-harmonic degrees 80–180. The coarse-grid checkpoint satisfies the predeclared shape and coherence margins in all four cells. The dense-grid checkpoint fails, with coherence changes of -0.038 for Q850 and -0.035 for U850 at query hour 2. The channels were selected during development, so this audit can promote a checkpoint but cannot confirm performance on an independently chosen spectral panel.

A final experiment isolates temporal reconstruction from forecast–analysis mismatch by taking both anchors and the target from one HRES run. A twelve-hour checkpoint reconstructs the archived six-hour state. On 16 initialisations, WeatherBridge gives the lowest point estimate among the three evaluated neural models, reducing RMSE by 42.6% relative to linear interpolation (Supplementary Table S2 and Supplementary Fig. S5). This small 2021 sample is exploratory and has no training-seed uncertainty.

Frozen-coefficient adapter

Before adaptation, Linear interpolation has the lowest point estimate in 93 of 120 field–hour cells at 6 h and 232 of 264 cells at 12 h. Within the displayed three-model

comparison, WeatherDCAE-14M and WeatherBridge lead in 16/11 cells at 6 h and 20/12 cells at 12 h, respectively. The twelve-hour transfer scores and raw six-hour cell counts remain in Supplementary Data 1, while Supplementary Fig. S12 places the raw six-hour controls and the fixed adapter on the 2022 test index. Nonlinear corrections learned from ERA5 anchor pairs therefore transfer poorly to HRES anchors. The validation-selected correction in Eq. 3 addresses this source shift. WeatherDCAE-14M and raw WeatherBridge are post-hoc controls; the confirmatory adapter-versus-Linear test excludes both post-hoc curves.

Across the 16-initialisation 2022 test, the adapter reduces aggregate RMSE from 0.73466 to 0.73382, a 0.114% change. None of 10,000 sign-flip draws is at least as extreme as the observed effect (Monte Carlo resolution 1/10001); the paired MSE-effect 95% interval is $[0.00116, 0.00129]$. Sign flips and bootstrap draws operate over the 16 forecast initialisations; the 3200 field–hour records remain grouped within each initialisation. The reduction is 0.137% at the held-out hours $\tau \in \{2, 4\}$, again with no exceedance in 10,000 sign flips. It is concentrated at short HRES leads: 2.384% at 0–48 h and 0.301% at 48–120 h, followed by the predeclared exact tie at 120–240 h. Among the 72 active field–hour cells, 70 improve and two regress; the remaining 48 cells are exact guarded ties. The two regressions are T700 at $\tau = 2$ (-0.048%) and $\tau = 4$ (-0.028%). Thus aggregate and held-out-hour RMSE improve despite two small T700 regressions.

The cell-centred SHT analysis in Supplementary Table S18 separates coefficient alignment from amplitude error. The frozen coefficient adapter increases high-wavenumber coherence and signed cospectrum for both Q850 and U850 at $\tau \in \{2, 3\}$. No sign-flip exceedance occurs in the four field–hour cells; the two-test Holm family has resolution 2/10001. Energy and shape errors are mixed: U850 shape improves at $\tau = 2$, whereas three of four energy comparisons and both Q850 shape comparisons favour Linear. Thus the adapter improves alignment of small-scale coefficients but does not fully calibrate their amplitude.

2.3 Spectral analysis

Supplementary Fig. S11 examines two six-hour query times: $\tau = 2$ was excluded from training, whereas $\tau = 3$ was observed. We report power ratio and signed cospectrum for Linear interpolation, the four neural baselines, WeatherDCAE-14M and WeatherBridge. Power can match despite misplaced or anti-phase structure; a signed cospectrum near one also requires aligned coefficients. Band-integrated energy, shape and phase statistics use every valid 2020 date at the 00 and 12 UTC anchor starts, with paired calendar-week resamples. Specific humidity (Q850) and zonal wind (U850) at 850 hPa were chosen during development to show moisture and wind behaviour and are interpreted descriptively.

WeatherBridge has the highest signed cospectrum among the plotted methods in all four Q850 and U850 panels, while the power-ratio ordering varies with field, wavenumber and query hour. Power and cospectrum must be read together: similar amplitude can coexist with weaker phase alignment. The all-field spectral tests determine the broader comparison. Supplementary Fig. S10 shows the corresponding 12 h power ratios for Q850 and V850 at a trained hour ($\tau = 5$) and a held-out hour ($\tau = 6$). The figure shows the component curves. The scalar and vector spherical-harmonic-transform (SHT) summaries in Supplementary Tables S12 and S13 determine the spectral comparison.

2.4 Twelve-hour interval-length ablation

To test sensitivity to anchor spacing without redefining the primary task, we double the six-hour anchor gap. All 12 h variants use the same 24 channels, 2017–2019 training years and query coordinate $t = \tau/\Delta$. They train at $\tau \in \{1, 2, 3, 5, 7, 9, 10, 11\}$ and are evaluated at every interior hour. The held-out hours $\{4, 6, 8\}$ include the midpoint and test interpolation between sampled query times. The learned path between those discrete hours remains unobserved. Methods gives model capacities and update budgets.

Supplementary Figs. S3 and S4 report the 12 h result by channel class. WeatherBridge is evaluated at all eleven interior hours using the same normalisation as the other models. It obtains mean $\mathrm{RMSE}_{\mathrm{norm}} = 0.10812$ over the full grid and 0.13277 on the held-out hours, compared with 0.22265 and 0.27820, respectively, for Linear Interp. Its mean ACC is 0.98660 over all hours and 0.98110 on the held-out subset. The per-field panel cannot isolate the source of an error at a held-out hour; the spatial backbone, optimisation and capacity also change across models. Off-scale ModAFNO values are marked explicitly in the Supplementary panel.

Supplementary Table S1 reports the aggregate 12 h held-out-hour comparison for the 2020 development year. All learned models use the same nominal 0.5° data split, query-hour sampler, peak learning rate and evaluation windows; their architecture-specific update counts are reported in Supplementary Table S16.

The full-year spherical-harmonic audit is reported by channel in Supplementary Table S13. The corresponding 6 h results appear in Supplementary Table S12; both tables use the block-average latitude coordinates with spherical strip-area quadrature and retain every valid date on the 00/12 UTC start-time index. At 12 h, WeatherBridge preserves less mean high-frequency energy than WeatherDCAE-14M (0.768 versus 0.793) and is farthest from unity for T700 and T2m. The advantage observed in selected six-hour phase diagnostics therefore does not extend uniformly to the longer interval.

2.5 Diagnostic physical balances

Pointwise accuracy does not guarantee that a reconstructed state has the same diagnostic balance as ERA5. Supplementary Table S14 reports the predicted-to-ERA5 ratio for the ageostrophic-wind RMSE and the hydrostatic residual. Each entry averages the interior query hours and 180 anchor windows selected uniformly from a 48-hour candidate grid in 2020. WeatherBridge is shown beside WeatherDCAE-14M and linear interpolation. A ratio of one reproduces the ERA5 residual scale. Values below one may

result from smoothing, so neither the minimum nor the maximum ratio is preferred; distance from one is the relevant quantity. This smaller balance audit is descriptive; the separate full-year physical suite in Methods supplies the paired robustness analysis.

2.6 Regional diurnal-cycle reconstruction

Global RMSE leaves regional timing and amplitude unresolved. We therefore aggregate T2m, U10 and V10 over four predefined regions and 12 monthly local-time climatologies. Supplementary Table S15 gives equal weight to each of the resulting 144 region–month–field strata and compares the same three methods as the balance audit. The amplitude statistic measures departure from the ERA5 climatological range, while the circular peak error measures local-time displacement. These descriptive regions expose local behaviour that global RMSE misses; they were excluded from model selection and provide no population-level regional inference.

2.7 Event-scale evaluation under atmospheric extremes

Independent reports defined six events before error inspection: the Texas cold outbreak, Pacific Northwest heatwave, Hurricane Ida, and Typhoons Surigae, Chanthu and Rai. The evaluation manifest records their regions and diagnostic fields. Six cases are too few to estimate an extreme-event climatology. They do test whether the aggregate 2021 ordering survives in high-impact local states without fine-tuning or test-time adaptation. The aggregate summary below covers all six events. Fig. 6 shows three cyclone cases selected by the rule stored in the figure provenance.

WeatherBridge has lower event-aggregate RMSE in 6/6 cases. Its mean normalised RMSE is 0.087, compared with 0.097 for WeatherDCAE-14M and 0.144 for Linear Interpolation, a 10.5% reduction relative to WeatherDCAE-14M. Across the four tropical cyclones, the corresponding wind-speed RMSEs are 0.449 and 0.482 $\mathrm{m\,s^{-1}}$, respectively, a 6.8% reduction.

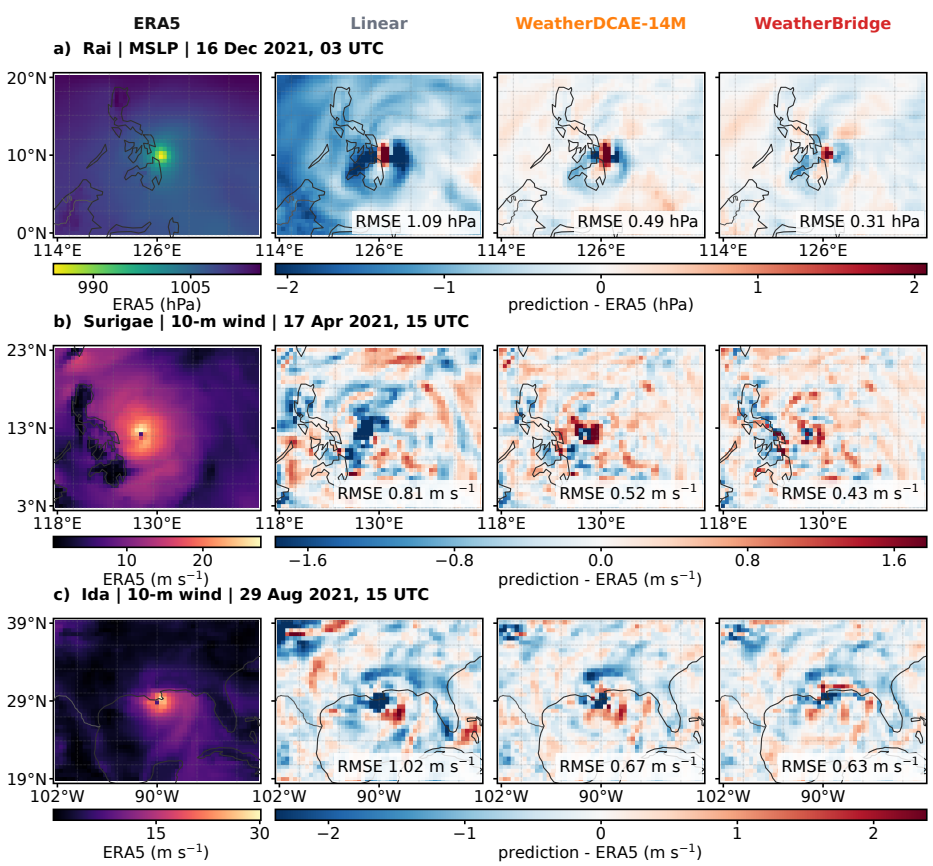


Fig. 6 Local diagnostics for three cases from the independently selected 2021 extreme-event set. In panels **a)**–**c)** the leftmost column is ERA5 and the remaining three columns are the signed error of Linear Interpolation, WeatherDCAE-14M and WeatherBridge at τ=3 h for Typhoon Rai mean sea-level pressure, Typhoon Surigae 10-m wind speed and Hurricane Ida 10-m wind speed. Error maps share a colour range within each row, and annotations give area-weighted RMSE over the displayed region. Maps are illustrative post-hoc diagnostics: from the four cyclone candidates they retain cases with lower WeatherBridge RMSE than WeatherDCAE-14M for the displayed field and hour. Aggregate results in the text retain all six events, including Chanthu.

WeatherDCAE-14M has the smaller maximum-wind error in three of the four cyclones and the smaller temperature-extremum error in both temperature cases. WeatherBridge has lower error over the surrounding event fields. The models therefore rank differently for distributed accuracy and isolated peak retention.

Supplementary Fig. S2 adds a development-year trajectory rather than a single selected hour. It shows mean sea-level pressure during Typhoon Haishen at three

consecutive query times, including both held-out hours. Supplementary Fig. S1 gives the corresponding vector-wind comparison during Hurricane Laura. Both are development-time qualitative diagnostics; the 2021 event audit excludes them.

3 Discussion

Learned corrections remain useful at sampled query hours omitted from optimisation. WeatherBridge retains its aggregate lead at the held-out hours for both six- and twelve-hour anchor gaps. On the 2022 ERA5 sample, held-out RMSE is 11.09% and 5.69% below WeatherDCAE-14M, respectively. The matched six-hour control quantifies the cost of withholding supervision: restoring hours 2 and 4 to training lowers their RMSE by about 8.8%, while changing trained-hour RMSE by less than 0.4%. The sparse-query model transfers useful skill, while direct supervision remains better. The evidence covers integer query hours; continuous-time evaluation and a matched twelve-hour control remain open.

WeatherBridge assigns different roles to transport and residual synthesis. The flow heads move resolved structure toward the requested hour; the local and global residual branches model deformation, unresolved development and smooth mass-field change. Removing acceleration from the trained checkpoint raises aggregate RMSE by 2.16%, and removing warped-anchor transport raises it by 60.54%. The hydrostatic head has a much smaller aggregate effect, but improves every geopotential/MSLP field–hour cell that it modifies and reduces mean hydrostatic-balance error. These internal displacements improve the prediction but represent neither observed parcel tracks nor conservation constraints.

Spectral power alone cannot distinguish a correctly placed front from one with the right amplitude at the wrong longitude. At six hours WeatherBridge improves signed cospectrum in several high-wavenumber diagnostics, while its power-ratio error depends on the field and query hour. At twelve hours WeatherDCAE-14M retains

more mean band energy, including clear advantages for T700 and T2m. The spectral loss therefore improves selected coefficient alignment without calibrating small-scale amplitude across both anchor gaps.

The IFS HRES experiments expose a second boundary. ERA5-trained corrections transfer poorly when both anchors come from a forecast trajectory. A coefficient adapter gives only a 0.114% aggregate gain, whereas full-weight adaptation across the deployed forecast-lead range gives 7.28% on the 24-initialisation 2022 audit. That number is retrospective because the full-range branch was designed after 2022 had been inspected. It also combines temporal reconstruction with correction of systematic forecast error; correlated endpoint and interior errors prevent an additive decomposition. An independent forecast archive is therefore needed before treating 7.28% as a prospective deployment estimate.

The physical and regional diagnostics add information that pointwise RMSE misses. Smoothing can reduce ageostrophic or hydrostatic residuals without matching the ERA5 balance, and four regional diurnal climatologies can expose local-time phase errors. The regions are descriptive examples. The separate robustness audit also records fieldwise, spectral, physical and temporal-curvature regressions. The resulting claim concerns aggregate interpolation at held-out hours with documented local regressions.

All models are trained against ERA5, whose hourly sequence contains assimilation increments as well as atmospheric evolution [27]. The 2021 audit changes the calendar year while retaining the data source, and the IFS HRES experiment covers a bounded set of forecast anchors. Native Pangu-Weather, GraphCast and other data-driven forecast trajectories remain untested because their grids, variable sets and source-specific errors require a separate transfer experiment. ERA5-relative accuracy may also reproduce the spatially varying observational constraints of the reanalysis.

The spatial conclusions apply to the stored nominal 0.5° grid. Evaluation uses the actual block-average strip areas, and the spectral diagnostics apply a spherical-harmonic transform at those coordinates. The transport trunk itself is a latitude–longitude network with circular padding in both tensor dimensions. Periodicity is correct in longitude but creates an artificial adjacency between the northernmost and southernmost rows. Only the low-wavenumber SFNO correction operates in a spherical-harmonic basis. A pole-aware trunk, such as a cubed-sphere convolutional network [49], would require matched retraining.

WeatherBridge is a deterministic 24-field interpolator for six-hour anchor gaps; the twelve-hour experiment tests sensitivity to a longer interval. Its evaluated scope ends at the later anchor and covers neither predictive uncertainty, precipitation nor 0.25° resolution. Architecture-specific paper-scale schedules support a quality comparison, not matched-compute or asymptotic rankings. The present checkpoint supports retrospective gap filling on a finer temporal grid; warnings, aviation routing and other safety-critical decisions lie outside its evaluated scope. Squared-error optimisation targets a conditional mean [50]; low RMSE is therefore insufficient evidence of realistic conditional variability. Evaluating a probabilistic interpolator would additionally require a sampling protocol and proper scores such as CRPS [51].

4 Methods

All learned models receive the same anchor states, static fields and query time, but differ in how they construct the interior state. We retain the endpoint chord explicitly so that each network learns transport or curvature relative to a common linear reference.

4.1 Problem formulation

Let $x_0, x_T \in \mathbb{R}^{C\times H\times W}$ denote the two anchor states with C=24 channels and (H, W)=(360, 720). With $t = \tau/\Delta \in (0, 1)$, the additive encoder–decoder baselines use

$$\hat{x}_\tau = \underbrace{(1-t)x_0 + t x_T}_{x_{\text{lin}}} + \tanh(\alpha)\, v_\theta(x_0, x_T, t, S), \tag{1}$$

where $\Delta \in \{6, 12\}$ h and the static-feature stack S contains the land-sea mask, normalised surface orography and the pinned cosine-latitude channel described in Sec. 4.4. The residual scale is initialised at 0.10, is scalar in WeatherDCAE and channelwise in PixelAttn-VFI, and has no post-activation floor. WeatherBridge uses the distinct transport-and-fusion path below; its residual gain is also channelwise and initialised at 0.10. All three retain x_{lin} as an explicit fallback. This separates the known endpoint chord from the temporal curvature and transported sub-anchor structure that must be learned.

4.2 Architectures evaluated

We evaluate the learned architecture families at 0.5° resolution with common anchor states, targets, query hours and scoring code. The main 6 h comparison uses six training years for every learned model, including WeatherDCAE-14M. The 12 h interval-length ablation uses the common three-year data protocol defined in Sec. 2.4. Learning rates, training years and update budgets are listed in Supplementary Table S16; architecture settings are in Supplementary Table S17.

Linear interpolation is the closed-form, zero-parameter reference and the fallback state used by the learned synthesis paths.

SwinV2.

The compact patch-token baseline concatenates the anchors and static fields into $[x_0, x_T, S]$. A 4×4 stride-four convolution produces 256-dimensional tokens. Eight SwinV2 blocks [46], arranged as four windowed/shifted-window pairs, use eight attention heads, a 5×9 latitude–longitude window, an MLP expansion ratio of four and stochastic depth increasing from zero to 0.1. A 128-component sinusoidal encoding of t is projected to featurewise scale and shift after every block. LayerNorm, a 1×1 projection, PixelShuffle and bilinear resizing return a 24-channel residual, which is multiplied by a learned scalar before addition to x_{lin}. The model borrows the SwinV2 block family used by FuXi [3], but not FuXi's cascade or full U-Transformer. The shifted-window hierarchy follows Swin [52]; the time-dependent scale and shift use featurewise conditioning [53].

ModAFNO and S-DYff.

ModAFNO [20, 54] uses a compact 36.9 M implementation at 6 h and the 151.7 M official `physicsnemo` configuration at 12 h; Supplementary Table S17 records both. The Fourier-operator lineage runs from FNO [55] to AFNO and its query-conditioned modulation. *S-DYff* is our interpolation-specific adaptation of DYffusion [21], using Spherical FNO [47] backbones. An interpolator supplies the initial state; a second network refines it for K=5 passes at the requested hour. This fixed-query refiner replaces the original forecaster and its diffusion-time schedule; it is not the published Spherical DYffusion climate emulator [23].

Neither DYffusion nor ModAFNO was validated here by reproducing the original paper's scores on its original data with an author checkpoint. Both were trained for the present field set and query protocol. Using the official ModAFNO configuration at 12 h establishes implementation provenance, not numerical reproduction of the published benchmark.

The original model rows use one checkpoint and one output per input, with no averaging across checkpoints, random seeds or stochastic draws. ModAFNO and the four direct/transport models run in evaluation mode without sampling. S-DYff also runs in evaluation mode, which disables ordinary dropout, but its custom stochastic-depth operation remains active with drop probability 0.1. Its row therefore measures a single stochastic reconstruction, not an ensemble mean. Stochastic depth [56] and Monte Carlo dropout [57] are distinct inference mechanisms. Original DYffusion evaluates ensemble-mean error and probabilistic scores; we do not reproduce that sampling protocol or report CRPS. The S-DYff ranking concerns this adaptation, not DYffusion's ensemble skill. The additional S-DYff-ENS curves average $N = 21$ stochastic reconstructions from the same checkpoint before computing RMSE and ACC. They retain the same stochastic-depth sampler, with ordinary dropout disabled; no model is retrained. S-DYff denotes the original $N = 1$ result throughout. To isolate the averaging effect, a separate paired comparison uses nested prefixes of $N = 1, 4, 16, 21$ draws from one sampling stream. Its $N = 1$ control is a fresh draw, not the archived S-DYff result. Seven-day block-bootstrap intervals quantify evaluation-date variability conditional on that sampling seed.

PixelAttn-VFI.

The two endpoints remain separate until the bottleneck. A shared encoder processes $[x_0, S]$ and $[x_T, S]$ through four scales with 72, 144, 288 and 576 channels; each scale uses two 3×3 convolution–GroupNorm–SiLU blocks, and the final three scales begin with stride-two downsampling. The deepest endpoint maps are flattened, concatenated and updated by eight-head self-attention [58]. Their two token sequences are then averaged and reshaped. A 128-component time embedding supplies a zero-initialised adaptive gate, following the conditioning idea in DiT [59], at this bottleneck. Three decoder stages apply a 4×4 stride-two transposed convolution, concatenate the mean endpoint features at the corresponding scale and apply the same two-convolution

block. A zero-initialised head predicts 24 residual channels with learned channelwise gains. The model tests feature-level endpoint interaction without a displacement field, correlation pyramid or warping operator.

WeatherDCAE-14M.

The direct-synthesis reference encodes $[x_0, x_T, S]$ through three 64/128/256-channel stages with three blocks per stage. The first two stages use residual convolutions [60]; the deepest uses three EfficientViT blocks [61] with multiscale linear attention, kernel scale five and 32-channel heads. PixelUnshuffle performs downsampling and a pole-aware projection produces a 256-channel latent. Longitude is padded periodically; samples crossing a pole are reflected in latitude and shifted by half the longitude span. Eight latitude rows are replicate-padded before the encoder to make the reported 0.5° checkpoint shape-compatible. The decoder uses two PixelShuffle stages and pole-aware residual blocks, but no cross-scale encoder–decoder skips. A final convolution predicts a 24-channel residual scaled by one learned scalar. This boundary operator is an equirectangular convention, not a rotationally equivariant spherical convolution.

Supplementary Fig. S13 traces these three comparator implementations from their common inputs to the residual output.

WeatherBridge.

The proposed transport-aware model uses $t = \tau/\Delta$ as query time normalised by the anchor interval. Its input is $[x_0, x_T, x_T - x_0, S]$, where S contains land–sea mask, orography and the pinned cosine-latitude feature. A sinusoidal embedding followed by a 256-dimensional multilayer perceptron conditions both the AdaLN-Zero bottleneck and the three decoder skip connections. The cross-scale encoder–decoder connections follow the U-Net pattern [62]. The shared block $B(c_{\text{in}}, c_{\text{out}}, s)$ applies a 3×3 convolution with stride s and the checkpoint's circular padding in both tensor dimensions, followed by GroupNorm [63] and SiLU; a second 3×3 convolution uses stride one. The longitude

wrap is physical, whereas the resulting adjacency between the northernmost and southernmost rows is an implementation artefact. For decoder level i, u is the upsampled decoder feature and E_i is the encoder feature at the same resolution. The block labelled "Skip connection" in Supplementary Fig. S14 computes $u + \tanh(g_i(t)) \odot P_{3\times3}[u, E_i]$ before applying B. Thus the short label refers to a query-time-gated skip rather than a plain U-Net concatenation. Here brackets denote channel concatenation, $P_{3\times3}$ is a learned convolution followed by GroupNorm and SiLU, and $g_i(t)$ is a channelwise gate obtained by a SiLU–linear projection of the time embedding. The symbol $\odot$ denotes elementwise multiplication.

Motion and anchor blending. The decoder produces $h_t^{(0)}$ at 360×720 and $h_t^{(1)}$ at 180×360, with 72 and 144 feature channels, respectively. Φ_{mot} and Φ_α denote two separate learned 3×3 convolutions applied to $h_t^{(0)}$. The motion head has eight outputs, arranged as four two-component maps: $(F_0, F_T, A_0, A_T) = 8\tanh(\Phi_{\mathrm{mot}}(h_t^{(0)}))$. F_0, F_T are first-order displacement coefficients and A_0, A_T are second-order coefficients, all in grid-pixel units for normalised time t. They are shared across weather channels and are not estimates of physical wind or material acceleration. The blend head has 24 outputs and gives $\alpha_t = \sigma(\Phi_\alpha(h_t^{(0)}) + 1 - 2t)$, where σ is the logistic sigmoid. Each channel and grid point therefore has its own weight for the two warped anchors. The offset $1 - 2t$ favours the nearer anchor before learning modifies the blend.

Additive reconstruction. Two further 3×3 convolutions map $h_t^{(0)}$ and $h_t^{(1)}$ to 24-channel fields Δ_t^{fine} and $\Delta_t^{\mathrm{coarse}}$. The operator up bilinearly resizes the coarse field to 360×720. The fine branch can change local amplitudes and shapes that displacement alone cannot reconstruct; the coarse branch predicts corrections with a wider spatial context. Both outputs are in normalised weather-field units, not displacement units. The third contribution, G_t^{SFNO}, comes from a spherical Fourier neural operator branch [47] applied to $h_t^{(0)}$. A 1×1 projection reduces its width to 20 channels; learned complex weights mix channels over the retained spherical-harmonic modes ($\ell, m < 20$). The

inverse transform is added to the projected features, followed by GELU and a 1×1 projection to 24 fields. This gives the residual global spatial context. Only its spectral operation is low-wavenumber: the local path and nonlinearity mean that the output need not be strictly band-limited. All three output projections are zero-initialised. Their sum is Δ_t; these are architectural corrections, not additional loss terms.

The two anchors are warped and combined with these corrections as follows:

$$
\begin{aligned}
d_0(t) &= F_0(t)t + \tfrac{1}{2}A_0(t)t^2, \qquad d_T(t) = F_T(t)(1-t) + \tfrac{1}{2}A_T(t)(1-t)^2, \\
W_0 &= \mathcal{W}(x_0, d_0), \qquad W_T = \mathcal{W}(x_T, d_T), \\
x_{\text{lin}} &= (1-t)x_0 + t x_T, \\
x_{\text{tr}} &= \alpha_t \odot W_0 + (1-\alpha_t) \odot W_T, \\
\Delta_t &= \Delta_t^{\text{fine}} + \text{up}(\Delta_t^{\text{coarse}}) + G_t^{\text{SFNO}}, \\
x_{\text{mix}} &= \beta \odot x_{\text{tr}} + (1-\beta) \odot x_{\text{lin}}, \\
x_{\text{pre}} &= x_{\text{mix}} + \tanh(\gamma) \odot \Delta_t, \\
\widehat{x}_t &= x_{\text{pre}} + \mathcal{I}_{Z,\text{MSLP}}\, H_T(x_{\text{pre}}),
\end{aligned}
\tag{2}
$$

where $\mathcal{W}$ is differentiable bilinear warping [64] on the lat–lon grid, $\beta = \sigma(b)$ is a learned per-channel gate between transport and the linear scaffold, γ is the learned per-channel residual gain. Unlike the spatially varying α_t, β and γ are learned vectors with one value per weather channel. x_{tr} is the transported-anchor blend, x_{mix} also includes the unwarped linear estimate, and x_{pre} adds the scaled residual.

Temperature-column correction. H_T is a zero-initialised 1×1 convolution that reads the four predicted temperature channels of x_{pre} at 1000, 925, 850 and 700 hPa and outputs five corrections. $\mathcal{I}_{Z,\text{MSLP}}$ inserts them into geopotential at those levels and MSLP, leaving the other 19 channels unchanged. In Supplementary Fig. S14, Δ_{hydro} denotes this 24-channel insertion, including its zero entries. The head is motivated by the relation between temperature and geopotential thickness. It learns a local column

adjustment; it does not integrate the hydrostatic equation or enforce an exact balance. The linear scaffold remains an explicit fallback while the model separately represents displacement, additive field changes and temperature–mass-field coupling. The trunk is a latitude–longitude convolutional network. We account for spherical geometry in evaluation through strip-area weights and spherical-harmonic diagnostics. Within the network, the spectral operation in the SFNO branch uses a spherical-harmonic basis. The convolutional encoder processes the anchors and static fields without time modulation. Conditioning enters at the AdaLN-Zero bottleneck and the decoder skip gates, making $F_i(t)$ and $A_i(t)$ query-specific coefficients. Each query receives its own second-order displacement; these coefficients describe separate query states rather than a parcel trajectory through all interior times.

WeatherDCAE-14M and WeatherBridge differ in both backbone and synthesis path; their comparison is architectural rather than a single-component ablation.

Retrained architecture variants.

Refine uses shared field controls, a 16-token area-weighted latent transformer and two decoder blocks per level, and omits the temperature-conditioned mass-field head. It has 13.7 million parameters and is trained from scratch. Several mechanisms change together, so this comparison does not isolate one component. Detail adds an anchor-detail bypass to the WeatherBridge trunk. The same flows and blend gate transport the finest pole-aware local multiband component of each anchor. A t-conditioned gain reinjects the difference between this transported detail and the model's own detail. The gain is initialised at zero, bounded to ± 0.5 and multiplied by $4t(1-t)$, where $t = \tau/\Delta$; Detail initially reproduces its parent. Supplementary Table S9 reports the retrained variants. The separate same-checkpoint lesions disable acceleration, warped-anchor transport or the hydrostatic head without retraining.

4.3 Training objective

The 14M-scale models minimise latitude-weighted pixel ℓ_1 error,

$$\mathcal{L}_{\mathrm{pix}} = \frac{1}{BCHW} \sum_{b,c,i,j} w_i \left| \hat{x}_{bcij} - x_{bcij} \right|,$$
$$w_i = \frac{H \cos(\widetilde{\phi}_i)}{\sum_{k=1}^{H} \cos(\widetilde{\phi}_k)}.$$

Here $\widetilde{\phi}_i$ denotes the legacy symmetric training rows $89.75°, \ldots, -89.75°$. Thus $\sum_i w_i = H$, and the displayed normalisation is exactly the tensor reduction implemented by the training code. The evaluated WeatherBridge run uses unit channel and query-time weights. This training convention records how the checkpoint was fitted. All reported evaluation reductions use the block-average coordinates in Sec. 4.4. The internal SFNO branch likewise uses the checkpointed `torch-harmonics` equiangular operator [47] on the stored tensor rows; it is a learned low-mode feature operator and is distinct from the double-precision diagnostic SHT. WeatherBridge adds two explicitly separated spectral terms:

$$\mathcal{L}_{\mathrm{WeatherBridge}} = \mathcal{L}_{\mathrm{pix}} + 0.05\mathcal{L}_{\mathrm{HP}} + 0.02\mathcal{L}_{\mathrm{FFT}}^{\mathrm{adv}}.$$

Here $\mathcal{L}_{\mathrm{HP}}$ is a phase-sensitive 3×3 local high-pass ℓ_1 loss with periodic-longitude and replicated-latitude padding. $\mathcal{L}_{\mathrm{FFT}}^{\mathrm{adv}}$ compares magnitudes of the planar fast Fourier transform (FFT) for the advected-channel mask. The training objective targets planar local detail and FFT amplitude, whereas the spherical-harmonic evaluation band $\ell \in [80, 180]$ is assessed independently. The spherical-harmonic-transform (SHT) diagnostic is computed independently after training. All 24 channels are scored; the FFT term is restricted to the declared advected-channel mask. WeatherDCAE-14M and PixelAttn-VFI use $\mathcal{L}_{\mathrm{pix}}$ alone. These 14M-scale runs use AdamW [65] with default moment coefficients, weight decay 10^{-4}, a 500-update linear warm-up followed by cosine decay, and gradient clipping at $\|g\|_2 = 1.0$.

SwinV2, ModAFNO, and S-DYff use a common forecasting-style baseline recipe. Their objective combines three terms: $\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anchor}} + \lambda_{\text{unit}}\mathcal{L}_{\text{unit}}$, where $\mathcal{L}_{\text{recon}}$ is a lat-weighted pixel-wise *smooth-*ℓ_1 (Huber) loss with β=0.02 and Aurora-style channel weights [4]. At 1000/925/850/700 hPa these weights are T: 1/1/1/1, U and V: 0.5/0.7/0.9/1.2, Q: 1.5/1.3/1.0/0.6, and Z: 1.0/0.8/0.5/0.3; the surface weights are 3.0/0.77/0.66/1.5 for T2m/U10/V10/MSLP. Smooth-ℓ_1 is quadratic for $|r| < \beta$ and linear in the tails. $\mathcal{L}_{\text{anchor}}$ penalises drift at τ=0, Δ (λ_{anc}=0.5, applied every four mini-batches). The residual branch predicts a correction v_θ to the linear interpolation x_{lin}. In SwinV2 and ModAFNO, the principal reconstruction loss scores $x_{\text{lin}} + g \odot v_\theta$, where g is the learned residual gain. Auxiliary unit-gain reconstruction evaluates the same correction with unit residual gain:

$$\mathcal{L}_{\text{unit}} = \mathcal{L}_{\text{recon}}(x_{\text{lin}} + v_\theta, x_\tau), \qquad \lambda_{\text{unit}} = 1.0.$$

It directly supervises the unscaled correction, even when g is small. For S-DYff, v_θ is the interpolator output before gain scaling, whereas the principal reconstruction loss scores the refined prediction. The forward pass separately preserves the sign of the learned gain and clamps its magnitude to at least 0.05. WeatherBridge also uses a linear interpolation and a residual branch, but trains that branch through the losses on its final output listed above; it does not use this auxiliary unit-gain reconstruction loss. Neither WeatherDCAE-14M nor PixelAttn-VFI uses $\mathcal{L}_{\text{unit}}$. SwinV2, ModAFNO, and S-DYff use AdamW ($\beta = (0.9, 0.95)$, weight decay 10^{-5}), cosine learning-rate decay, a three-epoch warm-up, and gradient clipping at $\|g\|_2 = 0.6$. Peak LR is 10^{-4} for every predicted channel and query hour in WeatherBridge, WeatherDCAE, SwinV2, ModAFNO, S-DYff, and PixelAttn-VFI. The 12 h held-out query-hour variants keep the same rate. Full budgets for the reported checkpoints are listed in Supplementary Table S16.

4.4 Data and evaluation protocol

We use ERA5 [16] through WeatherBench-2 [15]. The native 0.25° array (721×1440) is reduced to a nominal 0.5°, 360×720 array by averaging adjacent 2×2 samples after dropping the final −90° row. Consequently, the stored row midpoints are $89.875°, \ldots, -89.625°$ and the longitude midpoints are $0.125°, \ldots, 359.625°$, differing from the previously assumed symmetric half-cell grid. The main comparison scores 24 prognostic channels: 20 pressure-level fields (T, U, V, Q, Z at 1000/925/850/700 hPa) and four surface fields (T2m, U10, V10, MSLP). Every reported model uses this channel set for both anchor inputs and scored outputs. The chosen levels span the lower troposphere, from 1000–925 hPa near the surface to 850–700 hPa in the lower free troposphere. This scope keeps the global six-model benchmark tractable and limits its claims to lower-tropospheric interpolation. For the 6 h evaluation, train years are 2014–2019 (26,274 window–query training examples under the sparse-τ sampler), with training at $\tau \in \{1, 3, 5\}$ and evaluation at all five interior hours. The 12 h evaluation uses the 2017–2019 held-out query-hour protocol in Sec. 2.4. Calendar year 2020 supplies validation, architecture selection and the dense comparison (1463 windows per interior hour for 6 h); 2021 supplies the temporal-transfer audit. For the 2022 calendar check, we chose four dates per month (days 1, 8, 15 and 22), one 00 UTC anchor and 48 disjoint windows before generating predictions or inspecting target-derived scores. The full-year six-hour index contains anchor starts at 00, 06, 12 and 18 UTC. ERA5 uses 4D-Var windows from 09–21 and 21–09 UTC [66], so the 06 and 18 UTC pairs cross a window boundary, while the 00 and 12 UTC pairs remain within one window. We report these four strata separately in Supplementary Table S6. All runs read the same released per-channel normalisation files. Project records associate them with the 1990–2019 WeatherBench-2 ERA5 climatology at 00, 06, 12 and 18 UTC, cosine-latitude spatial weighting and a factor 1.3 on climatological standard deviations; the original generator was not retained.

An independent reconstruction matched pressure-level means to a maximum relative difference of 2.46×10^{-7}, but was not bit-exact. Its largest relative standard-deviation difference was 2.31×10^{-5} (T925), and the rounded MSLP standard deviation differed by 0.3 Pa. We therefore publish the original files and channel order as the canonical scale. Baselines using channel reweighting apply the Aurora-style loss weights [4].

The static-feature artifact was generated before this coordinate audit and is kept unchanged to preserve the inputs used by every checkpoint. Its third channel is $\cos\tilde{\phi}$ on the legacy symmetric rows $\tilde{\phi} = 89.75°, \ldots, -89.75°$, a uniform $-0.125°$ offset from the block-average row centres. The maximum absolute difference from cosine latitude evaluated at the field centres is 2.18×10^{-3}. All area weights, ACC reductions, physical diagnostics and spherical-harmonic transforms use the actual block-average coordinates above.

The paper-reported runs use A100-80GB GPUs; Supplementary Table S16 lists the epoch and update budgets. We report per-channel latitude-area-weighted RMSE normalised by the training standard deviation, ACC against the WeatherBench-2 1990–2019 hourly climatology, and spherical-harmonic diagnostics. All reductions use the actual block-average coordinates above. The climatology is block-averaged from the same native-row pairs as ERA5. The release includes its coordinates, source metadata and conversion code.

The displayed ACC pools anomaly cross-products and norms over all windows and grid cells for each field and hour. Paired tests instead compute spatial ACC per window and field before averaging over hours, which preserves the calendar-block sampling unit. The pooled curve and paired test therefore use different estimators and have different sampling intervals.

The spectral diagnostic uses a double-precision real SHT evaluated at the block-average rows with spherical strip-area weights. Standardised fields are multiplied by their training standard deviations, without restoring the constant mean, before scalar

and vector transforms. We retain $\ell \in [80, 180]$, or angular wavelengths of approximately 2.0°–4.5°, to test smoothing near the resolved small scales. Kinetic-energy spectra are also used to assess the effective resolution of numerical weather models [67]. This band is not fitted to the atmospheric regimes reported from observations [68, 69]. The diagnostics separate band energy, energy-normalised spectral shape, magnitude-squared coherence and signed normalised cospectrum; the vector SHT further separates spheroidal and toroidal wind energy. Every valid 2020 anchor day is sampled at two UTC times. Intervals use B=10,000 paired draws over calendar-week bins, with Holm–Bonferroni correction over all query hours and channels within each scalar or vector metric family. Strip areas are bounded by row midpoints and the poles. We restore negative longitudinal modes, clip energy ratios to $[10^{-8}, 10^{8}]$ before computing log error, and floor energy-normalised shapes at 10^{-12}.

Calendar-transfer sample and robustness audit.

After development on 2020 and inspection of the 2021 transfer year, we froze the candidate checkpoints and a 48-window 2022 ERA5 index before generating predictions on that index. We report aggregate and held-out RMSE with paired seven-day block intervals. A separate internal robustness rule asked the more conservative question of whether any fieldwise RMSE, hard-window, ACC, spectral, physical or temporal-curvature regression could be detected after familywise correction. It also checked three paired IFS HRES seeds. Section 4.8 defines the rule; Supplementary Table S7 and Supplementary Data 1 report its outcomes. HRES fields use the same 2×2 block average as ERA5, with source, coordinate and missing-value provenance stored in each artifact.

Low-cost forecast-anchor adaptation.

The raw forecast-anchor results motivate a coefficient correction for the ERA5-trained interpolator. We keep WeatherBridge fixed and write its prediction as $M_{\tau,c}$ and the

analytic interpolation as $L_{\tau,c}$. The adapted output is

$$\widehat{x}_{\tau,c} = L_{\tau,c} + s_r \left[g_{\tau,c}(M_{\tau,c} - L_{\tau,c}) + b_{\tau,c}\sigma_c \right], \tag{3}$$

where c indexes fields, σ_c is the ERA5 training standard deviation, and s_r is a scalar for the HRES lead range r. The constrained $g_{\tau,c} \in [0, 1]$ and $b_{\tau,c} \in [-0.25, 0.25]$ are computed directly from streaming second moments. Six initialisations from 2017–2019 fit the coefficients at $\tau \in \{1, 3, 5\}$; held hours are linearly interpolated in query time. Two 2020 initialisations select global shrinkage and the lead-range scales. A correction is disabled when its validation RMSE gain is below 0.1%.

On the declared 2021 development split, all remaining field–hour regressions were confined to $\tau \in \{1, 5\}$ and were below 0.01%. We therefore selected exact Linear output at those two query hours before opening 2022. The same guard selected exact Linear for HRES leads 120–240 h; the learned correction is used only at $\tau \in \{2, 3, 4\}$ for leads below 120 h. The JSON adapter stores 240 field–hour coefficients, of which 144 remain active, and three lead scales in 27 kB. At inference the guarded branches bypass the neural model, reducing the number of network query evaluations by 70% on the lead grid. The 2022 test uses 16 prespecified initialisations with unchanged adapter parameters. This adapter belongs to the primary 6 h task; the 12 h interval-length ablation excludes it.

Full-weight HRES adaptation.

We also update every WeatherBridge parameter using IFS HRES anchor pairs and ERA5 at the interior valid time. Optimisation uses 108 forecast trajectories initialised in 2017–2019 at query hours 1, 3 and 5. Twenty-four 2020 trajectories and all five interior hours select the checkpoint. The full-range run samples left-anchor leads from 0 to 216 h on a 24 h grid, trains for eight epochs with AdamW at peak learning rate 10^{-5}, and retains the ERA5 high-pass and advected-spectrum losses. The 2022 audit

uses 24 initialisations on the fifth and nineteenth day of each month, left-anchor leads $\{0, 24, 48, 72, 96\}$ h and targets through 102 h. These dates predate the full-range run, but 2022 had already been inspected during HRES development. We therefore treat the result as retrospective. The public WeatherBench-2 HRES archive ends in 2022, so confirmation requires a later operational archive or an independently sourced forecast trajectory.

4.5 Implementation and compute

Runs use PyTorch 2.5.1, CUDA 12.4, PyTorch Lightning 2.6.1 and `torch-harmonics` 0.7.3 with mixed precision. The 12 h ModAFNO follows the official `physicsnemo` reference; the compact 6 h variant uses our implementation. Training and evaluation use two NVIDIA A100-80GB SXM4 GPUs and a 486 GiB host-memory allocation. Supplementary Tables S16 and S17 report the per-model training budgets, parameter counts and architecture settings.

Inference latency is the median of seven CUDA-event repeats, each containing 20 bfloat16 forward passes after five warm-ups, on one A100-80GB at batch size one and $24 \times 360 \times 720$ resolution. The benchmark uses random-input seed 2027, cycles $t \in \{0.25, 0.50, 0.75\}$ and compares every candidate with the fastest model measured in the same process.

4.6 Physical-consistency diagnostics

A full-year per-window diagnostic suite complements the smaller balance table in Results. It contains normalised MSE for horizontal wind divergence, relative vorticity and kinetic energy; MSE between predicted and ERA5 hydrostatic residuals; area-weighted negative specific-humidity fraction; absolute relative global MSLP bias; and absolute relative lower-tropospheric moisture bias. Divergence and vorticity use spherical finite differences and exclude latitudes poleward of 75°; all other reductions use the declared strip areas. Every quantity is non-negative and lower is better. These

values are stored for each full-year evaluation window and compared with paired seven-day blocks.

The smaller balance audit uses uniformly spaced windows and measures two residual magnitudes; the robustness audit evaluates all seven physical diagnostics. Geostrophic wind is derived from the horizontal geopotential gradient at 850 and 700 hPa using centred spherical finite differences. We then compute the latitude-weighted RMSE of the ageostrophic residual over 15°–70° absolute latitude. Hydrostatic consistency is measured on the adjacent 1000/925/850/700 hPa layers using $\Delta Z + R_d \overline{T_v} \Delta \ln p$, where $T_v = T(1 + 0.61q)$, and is normalised by the ERA5 residual from the same windows. Both diagnostics use the declared spherical strip areas of the stored block-average grid. These diagnostics are evaluated for both 6 h and 12 h anchor spacings at every interior query hour. We select 180 start times uniformly from the 2020 48-hour candidate grid and reuse the exact index for every model within a horizon. None of these quantities enters the training loss.

4.7 Diurnal-cycle diagnostic

We reconstruct hourly T2m, U10 and V10 sequences to test regional timing and amplitude independently of global RMSE. Diurnal amplitude and peak hour are evaluated in predefined 5° boxes over the Sahel, Amazon, Congo and south-eastern Australia. The diagnostic reports amplitude bias and circular peak-time error relative to ERA5 for all four compared methods. We average each error equally over four regions, 12 calendar months and three surface fields, giving 144 strata per model and horizon. The amplitude statistic is $|A_{\mathrm{recon}}/A_{\mathrm{ERA5}} - 1|$; peak-time error is $\min(|\Delta h|, 24 - |\Delta h|)$.

4.8 Reproducibility and statistical analysis

Linear interpolation and learned models use identical anchor pairs. In the principal architecture comparison, 95% intervals for aggregate RMSE use B=5000 paired bootstrap draws over contiguous seven-day blocks of the common full-year index. Block

resampling preserves local temporal dependence [70]; independent resampling of hourly fields would disregard that dependence. Calendar splits address a separate problem: leakage between temporally related training and evaluation data [71]. Pairing is retained within each draw, and the channel set and all query hours remain grouped with their anchor block. The full-year tests contain 53 calendar-week blocks. Sparse hard-window subsets retain 19–30 non-empty blocks, with the exact `n_blocks` deposited for each comparison; draws lacking any requested query hour are discarded and redrawn; the output artifact records the number discarded.

The six-hour full-query control uses the same 2014–2019 training pool, anchor construction, 24 fields and full-year 2020/2021 evaluation windows as the sparse-query runs. For WeatherBridge, seed, optimiser-step budget, effective batch size, precision and loss weights are matched; the training query set changes from $\{1, 3, 5\}$ to $\{1, 2, 3, 4, 5\}$ h. Its paired sparse-minus-full RMSE intervals use 5,000 day-block bootstrap draws, rather than the seven-day blocks of the principal comparison. They measure evaluation-period variability for one matched training seed. The other architectures have unequal or incompletely documented seeds or update budgets (Supplementary Table S4) and are reported as diagnostic comparisons.

Mixed-precision kernels and cuDNN operations prevent bit-exact reproduction. The intervals quantify test-period sampling variability. Training sensitivity is reported through three paired seed-to-seed forecast-anchor comparisons, which provide paired effects rather than a distributional seed-level interval.

Supplementary Data 1 gives the observed effect, 95% interval and block-swap permutation p value for every inferential comparison in the robustness audit. Aggregate tests are two-sided; cellwise improvement and regression tests are one-sided. The significance threshold is $\alpha = 0.05$ after correction within each family. Cellwise families retain both raw and Holm–Bonferroni-adjusted values [72], together with effect directions. Raw Monte Carlo p values use the plus-one estimator $(r + 1)/(B + 1)$. Values at

$1/(B+1)$ mean that no resampled statistic was as extreme as the observation; they are marked `p_censored=true` instead of being printed as exact tail probabilities. The study was not preregistered.

Field–hour RMSE, upper-5%-hard-window RMSE and ACC each form a separate family of $24(\Delta-1)$ tests for each candidate, horizon and year, with Δ expressed in hours. Temporal curvature forms a 24-field family; physical tests are corrected across the named diagnostics; each scalar and vector spectral metric is corrected across its query-hour–field grid. The 2020 and 2021 families are kept separate. The internal robustness sensitivity rejects a candidate with any Holm-significant regression in these families against WeatherDCAE-14M. It also requires at most 15 million parameters, latency no greater than 1.25 times the fastest candidate, non-increasing aggregate and held-out RMSE, and paired IFS HRES improvement in at least two of three seeds with a non-positive mean change. This rule was fixed after development on 2020 and inspection of 2021. Supplementary Information reports its outcomes separately from the primary aggregate comparison. This familywise accounting avoids interpreting isolated significant field–hour cells as independent confirmations [73].

4.9 Use of generative AI

Generative AI tools assisted with language editing and translation, manuscript revision, and preparation of code and figure layouts. Weather fields and reported metrics come from the datasets and model evaluations described in Methods. The authors are responsible for the code, analyses, figures, references and conclusions in the final manuscript.

Data availability. ERA5 is available from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/). The hourly ERA5 source used here is the public WeatherBench-2 Zarr store at gs://weatherbench2/datasets/era5/1959-2023_01_10-full_37-1h-0p25deg-chunk-1.zarr; its metadata can be opened without a Google account at https://storage.googleapis.com/weatherbench2/datasets/era5/1959-2023_01_10-full_37-1h-0p25deg-chunk-1.zarr/.zmetadata. The IFS HRES audit uses gs://weatherbench2/datasets/hres/2016-2022-0012-1440x721.zarr, with public metadata at https://storage.googleapis.com/weatherbench2/datasets/hres/2016-2022-0012-1440x721.zarr/.zmetadata. The data chunks are available without a Google account under the same store prefixes; access instructions are in the WeatherBench-2 data guide (https://weatherbench2.readthedocs.io/en/latest/data-guide.html). Methods specifies the 0.5° product, splits, normalisation and channel set. Supplementary Data 1 reports paired effects, intervals and p values. Supplementary Data 2 contains the prespecified 2022 WeatherBridge assessment together with the evaluation index and model manifest it was computed from.

Code availability. Source code, experiment configurations, figure-generation scripts and seven trained checkpoints are available at https://github.com/dsuhoi/weatherbridge (version 1.0.1, https://github.com/dsuhoi/weatherbridge/tree/v1.0.1). The versioned snapshot includes metric summaries, normalisation, static fields and an example ERA5 window; weights are distributed through Git LFS. It provides WeatherBridge at 6 h and 12 h intervals and the five learned 6 h comparators. Other 12 h comparator, full-query control and HRES-adapted checkpoints are not included. Full-year inference requires the external datasets described above; redrawing figures from stored scores does not rerun inference.

Funding. This study received no funding.

Author contributions. D.S. led the study, implemented the models and evaluation protocol, performed most of the experiments and analyses, prepared the main figures,

and wrote and revised the manuscript. A.Z. implemented and conducted the full-query-hour ablation experiments and prepared the corresponding figures. Y.M., I.M. and I.O. supervised the study, contributed to the development and discussion of the research ideas, and revised the manuscript. All authors discussed the results and approved the final manuscript.

Competing interests. The authors declare no competing interests.

References

[1] Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023). https://doi.org/10.1126/science.adi2336.

[2] Bi, K. *et al.* Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023). https://doi.org/10.1038/s41586-023-06185-3.

[3] Chen, L. *et al.* FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *NPJ Climate and Atmospheric Science* **6**, 190 (2023). https://doi.org/10.1038/s41612-023-00512-1.

[4] Bodnar, C. *et al.* A foundation model for the Earth system. *Nature* **641**, 1180–1187 (2025). https://doi.org/10.1038/s41586-025-09005-y.

[5] Nguyen, T. *et al. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. Advances in Neural Information Processing Systems*, Vol. 37, 68740–68771 (2024). https://doi.org/10.52202/079017-2196.

[6] Moldovan, G. *et al.* AIFS Single 1.1.0: an update to ECMWF's machine-learned weather forecast model AIFS. *Geoscientific Model Development* **19**, 4703–4724 (2026). https://doi.org/10.5194/gmd-19-4703-2026.

[7] Price, I. *et al.* Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025). https://doi.org/10.1038/s41586-024-08252-9.

[8] Kurth, T. *et al. FourCastNet: Accelerating global High-Resolution weather forecasting using adaptive Fourier neural operators. Proceedings of the Platform for Advanced Scientific Computing Conference*, 1–11 (2023). https://doi.org/10.1145/3592979.3593412.

[9] Kochkov, D. *et al.* Neural general circulation models for weather and climate. *Nature* **632**, 1060–1066 (2024). https://doi.org/10.1038/s41586-024-07744-y.

[10] Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K. & Grover, A. *ClimaX: A foundation model for weather and climate. Proceedings of the 40th International Conference on Machine Learning*, Vol. 202 of *Proceedings of Machine Learning Research*, 25904–25938 (2023). URL https://proceedings.mlr.press/v202/nguyen23a.html.

[11] Chen, K. *et al.* The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time. *Communications Earth & Environment* **6**, 518 (2025). https://doi.org/10.1038/s43247-025-02502-y.

[12] Allen, A. *et al.* End-to-end data-driven weather prediction. *Nature* **641**, 1172–1179 (2025). https://doi.org/10.1038/s41586-025-08897-0.

[13] Keisler, R. Forecasting global weather with graph neural networks. *arXiv preprint arXiv:2202.07575* (2022). https://doi.org/10.48550/arXiv.2202.07575.

[14] Rasp, S. *et al.* WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems* **12**, e2020MS002203 (2020). https://doi.org/10.1029/2020ms002203.

[15] Rasp, S. *et al.* WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems* **16**, e2023MS004019 (2024). https://doi.org/10.1029/2023ms004019.

[16] Hersbach, H. *et al.* The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society* **146**, 1999–2049 (2020). https://doi.org/10.1002/qj.3803.

[17] Bowler, N. E., Pierce, C. E. & Seed, A. Development of a precipitation nowcasting algorithm based upon optical flow techniques. *Journal of Hydrology* **288**, 74–91 (2004). https://doi.org/10.1016/j.jhydrol.2003.11.011.

[18] Pulkkinen, S. *et al.* Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development* **12**, 4185–4219 (2019). https://doi.org/10.5194/gmd-12-4185-2019.

[19] Staniforth, A. & Côté, J. Semi-lagrangian integration schemes for atmospheric models—a review. *Monthly Weather Review* **119**, 2206–2223 (1991). https://doi.org/10.1175/1520-0493(1991)119⟨2206:slisfa⟩2.0.co;2.

[20] Leinonen, J., Bonev, B., Kurth, T. & Cohen, Y. Modulated adaptive Fourier neural operators for temporal interpolation of weather forecasts. *arXiv preprint arXiv:2410.18904* (2024). https://doi.org/10.48550/arXiv.2410.18904.

[21] Rühling Cachay, S., Zhao, B., Joren, H. & Yu, R. *DYffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. Advances in Neural Information Processing Systems*, Vol. 36 (2023). URL https://papers.nips.cc/paper_files/paper/2023/hash/8df90a1440ce782d1f5607b7a38f2531-Abstract-Conference.html.

[22] Ho, J., Jain, A. & Abbeel, P. *Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems*, Vol. 33, 6840–6851 (2020). URL https://proceedings.neurips.cc/paper/2020/hash/

4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

[23] Rühling Cachay, S., Henn, B., Watt-Meyer, O., Bretherton, C. S. & Yu, R. *Probabilistic emulation of a global climate model with Spherical DYffusion. Advances in Neural Information Processing Systems*, Vol. 37 (2024). https://doi.org/10.52202/079017-4052.

[24] Glawion, L., Polz, J., Kunstmann, H., Fersch, B. & Chwala, C. Global spatio-temporal ERA5 precipitation downscaling to km and sub-hourly scale using generative AI. *NPJ Climate and Atmospheric Science* **8**, 219 (2025). https://doi.org/10.1038/s41612-025-01103-y.

[25] Wang, J. *et al.* Neural network based on high freedom dynamic collaboration of flows for meteorological data temporal downscaling. *ISPRS Journal of Photogrammetry and Remote Sensing* **234**, 170–184 (2026). https://doi.org/10.1016/j.isprsjprs.2026.02.017.

[26] Gao, S., Lin, L., Zhang, Z. & Wang, J. Temporal downscaling meteorological variables to unseen moments: Continuous temporal downscaling via multi-source spatial–temporal-wavelet feature fusion and time-continuous manifold. *ISPRS Journal of Photogrammetry and Remote Sensing* **230**, 32–54 (2025). https://doi.org/10.1016/j.isprsjprs.2025.09.001.

[27] Ingstad, M. S. *et al.* Hourglass: A probabilistic data-driven temporal downscaler for global and regional weather forecasting. *arXiv preprint arXiv:2607.11457* (2026). https://doi.org/10.48550/arXiv.2607.11457.

[28] Ravuri, S. *et al.* Skilful precipitation nowcasting using deep generative models of radar. *Nature* **597**, 672–677 (2021). https://doi.org/10.1038/s41586-021-03854-z.

[29] Zhang, Y. *et al.* Skilful nowcasting of extreme precipitation with NowcastNet. *Nature* **619**, 526–532 (2023). https://doi.org/10.1038/s41586-023-06184-4.

[30] Sønderby, C. K. *et al.* MetNet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140* (2020). https://doi.org/10.48550/arXiv.2003.12140.

[31] Mardani, M. *et al.* Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment* **6**, 124 (2025). https://doi.org/10.1038/s43247-025-02042-5.

[32] Kalluri, T., Pathak, D., Chandraker, M. & Tran, D. *FLAVR: Flow-agnostic video representations for fast frame interpolation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2071–2082 (2023). https://doi.org/10.1109/wacv56688.2023.00211.

[33] Shi, Z., XU, X., Liu, X., Chen, J. & Yang, M.-H. *Video frame interpolation transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17482–17491 (2022). https://doi.org/10.1109/cvpr52688.2022.01696.

[34] Zhang, G. *et al. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5682–5692 (2023). https://doi.org/10.1109/cvpr52729.2023.00550.

[35] Bao, W. *et al. Depth-aware video frame interpolation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3698–3707 (2019). https://doi.org/10.1109/CVPR.2019.00382.

[36] Huang, Z., Zhang, T., Heng, W., Shi, B. & Zhou, S. *Real-time intermediate flow estimation for video frame interpolation. Computer Vision – ECCV 2022*, Vol. 13674, 624–642 (2022). https://doi.org/10.1007/978-3-031-19781-9_36.

[37] Kong, L. *et al. IFRNet: Intermediate feature refine network for efficient frame interpolation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1969–1978 (2022). https://doi.org/10.1109/cvpr52688.2022.00201.

[38] Li, Z. *et al. AMT: All-pairs multi-field transforms for efficient frame interpolation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9801–9810 (2023). https://doi.org/10.1109/cvpr52729.2023.00945.

[39] Xu, X., Li, S., Sun, W., Yin, Q. & Yang, M.-H. *Quadratic video interpolation. Advances in Neural Information Processing Systems*, Vol. 32 (2019). URL https://proceedings.neurips.cc/paper/2019/hash/d045c59a90d7587d8d671b5f5aec4e7c-Abstract.html.

[40] Liu, Y. *et al. Enhanced quadratic video interpolation. Computer Vision – ECCV 2020 Workshops*, Vol. 12538, 41–56 (2020). https://doi.org/10.1007/978-3-030-66823-5_3.

[41] Niklaus, S. & Liu, F. *Softmax splatting for video frame interpolation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5437–5446 (2020). https://doi.org/10.1109/cvpr42600.2020.00548.

[42] Jiang, H. *et al. Super SloMo: High quality estimation of multiple intermediate frames for video interpolation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9000–9008 (2018). https://doi.org/10.1109/cvpr.2018.00938.

[43] Reda, F. *et al. FILM: Frame interpolation for large motion. Computer Vision – ECCV 2022*, Vol. 13667, 250–266 (2022). https://doi.org/10.1007/978-3-031-20071-7_15.

[44] Voleti, V., Jolicoeur-Martineau, A. & Pal, C. *MCVD - masked conditional video diffusion for prediction, generation, and interpolation. Advances in Neural Information Processing Systems*, Vol. 35, 23371–23385 (2022). https://doi.org/10.52202/068431-1698.

[45] Danier, D., Zhang, F. & Bull, D. *LDMVFI: Video frame interpolation with latent diffusion models. Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, 1472–1480 (2024). https://doi.org/10.1609/aaai.v38i2.27912.

[46] Liu, Z. *et al. Swin transformer V2: Scaling up capacity and resolution. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12009–12019 (2022). https://doi.org/10.1109/cvpr52688.2022.01170.

[47] Bonev, B. *et al. Spherical Fourier neural operators: Learning stable dynamics on the sphere. Proceedings of the 40th International Conference on Machine Learning*, Vol. 202 of *Proceedings of Machine Learning Research*, 2806–2823 (2023). URL https://proceedings.mlr.press/v202/bonev23a.html.

[48] Chen, J. *et al. Deep compression autoencoder for efficient High-Resolution diffusion models. The Thirteenth International Conference on Learning Representations* (2025). URL https://openreview.net/forum?id=wH8XXUOUZU.

[49] Weyn, J. A., Durran, D. R. & Caruana, R. Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *Journal of Advances in Modeling Earth Systems* **12**, e2020MS002109 (2020). https://doi.org/10.1029/2020MS002109.

[50] Gneiting, T. Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**, 746–762 (2011). https://doi.org/10.1198/jasa.2011.r10138.

[51] Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378 (2007). https://doi.org/10.1198/016214506000001437.

[52] Liu, Z. *et al. Swin transformer: Hierarchical vision transformer using shifted windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992–10002 (2021). https://doi.org/10.1109/iccv48922.2021.00986.

[53] Perez, E., Strub, F., de Vries, H., Dumoulin, V. & Courville, A. *FiLM: Visual reasoning with a general conditioning layer. Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32 (2018). https://doi.org/10.1609/aaai.v32i1.11671.

[54] Guibas, J. *et al. Adaptive Fourier neural operators: Efficient token mixers for transformers. International Conference on Learning Representations* (2022). URL https://openreview.net/forum?id=EXHG-A3jlM.

[55] Li, Z. *et al. Fourier neural operator for parametric partial differential equations. International Conference on Learning Representations* (2021). URL https://openreview.net/forum?id=c8P9NQVtmnO.

[56] Huang, G., Sun, Y., Liu, Z., Sedra, D. & Weinberger, K. Q. *Deep networks with stochastic depth. Computer Vision – ECCV 2016*, Vol. 9908, 646–661 (2016). https://doi.org/10.1007/978-3-319-46493-0_39.

[57] Gal, Y. & Ghahramani, Z. *Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning*, Vol. 48 of *Proceedings of Machine Learning Research*,

1050–1059 (2016). URL https://proceedings.mlr.press/v48/gal16.html.

[58] Vaswani, A. *et al. Attention is all you need. Advances in Neural Information Processing Systems*, Vol. 30 (2017). URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[59] Peebles, W. & Xie, S. *Scalable diffusion models with transformers. 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4172–4182 (2023). https://doi.org/10.1109/ICCV51070.2023.00387.

[60] He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016). https://doi.org/10.1109/cvpr.2016.90.

[61] Cai, H., Li, J., Hu, M., Gan, C. & Han, S. *EfficientViT: Lightweight multi-scale attention for High-Resolution dense prediction. 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 17256–17267 (2023). https://doi.org/10.1109/iccv51070.2023.01587.

[62] Ronneberger, O., Fischer, P. & Brox, T. *U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Vol. 9351, 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28.

[63] Wu, Y. & He, K. *Group normalization. Computer Vision – ECCV 2018*, Vol. 11217, 3–19 (2018). https://doi.org/10.1007/978-3-030-01261-8_1.

[64] Jaderberg, M., Simonyan, K., Zisserman, A. & Kavukcuoglu, K. *Spatial transformer networks. Advances in Neural Information Processing Systems*, Vol. 28 (2015). URL https://proceedings.neurips.cc/paper/2015/hash/33ceb07bf4eeb3da587e268d663aba1a-Abstract.html.

[65] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization. International Conference on Learning Representations* (2019). URL https://openreview.net/forum?id=Bkg6RiCqY7.

[66] European Centre for Medium-Range Weather Forecasts. ERA5 data documentation. Copernicus Knowledge Base (2026). URL https://confluence.ecmwf.int/pages/viewpage.action?pageId=348786747. Accessed 20 August 2026.

[67] Skamarock, W. C. Evaluating mesoscale NWP models using kinetic energy spectra. *Monthly Weather Review* **132**, 3019–3032 (2004). https://doi.org/10.1175/MWR2830.1.

[68] Nastrom, G. D. & Gage, K. S. A climatology of atmospheric wavenumber spectra of wind and temperature observed by commercial aircraft. *Journal of the Atmospheric Sciences* **42**, 950–960 (1985). https://doi.org/10.1175/1520-0469(1985)042⟨0950:acoaws⟩2.0.co;2.

[69] LINDBORG, E. Can the atmospheric kinetic energy spectrum be explained by two-dimensional turbulence? *Journal of Fluid Mechanics* **388**, 259–288 (1999). https://doi.org/10.1017/s0022112099004851.

[70] Künsch, H. R. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* **17**, 1217–1241 (1989). https://doi.org/10.1214/aos/1176347265.

[71] Roberts, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40**, 913–929 (2017). https://doi.org/10.1111/ecog.02881.

[72] Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70 (1979). URL https://www.jstor.org/stable/4615733.

[73] Wilks, D. S. The stippling shows statistically significant grid points: How research results are routinely overstated and overinterpreted, and what to do about it. *Bulletin of the American Meteorological Society* **97**, 2263–2273 (2016). https://doi.org/10.1175/bams-d-15-00267.1.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementary.pdf