Supplementary Information for 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;

1 Secondary cyclone diagnostic

Supplementary Fig. S1 compares the 10-m wind vector during Hurricane Laura. This development-time case lies outside the 2021 event audit. All learned rows use the checkpoints reported in the main comparison.

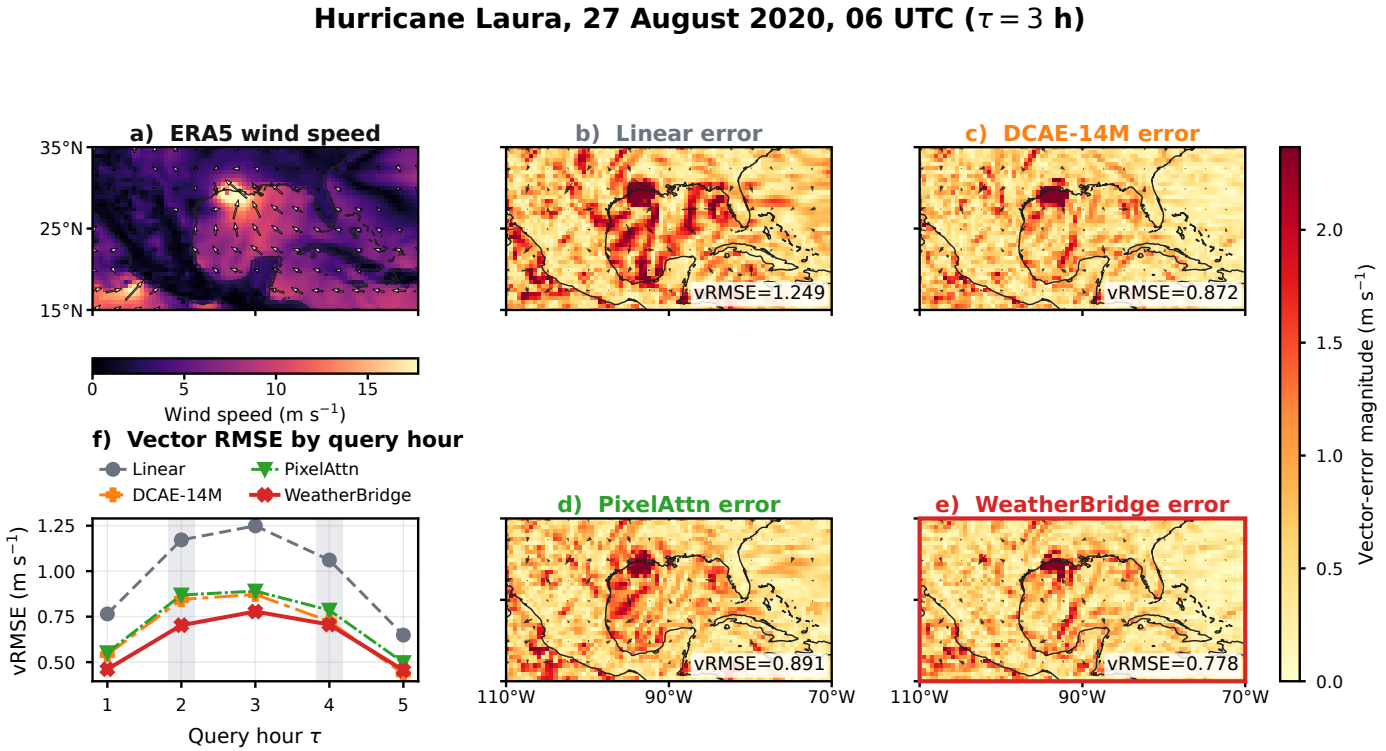


Fig. S1 10-m wind reconstruction during Hurricane Laura on 27 August 2020. The left column shows ERA5 wind speed and direction at $\tau = 3$ h and the five-hour vector-RMSE trajectories. The right 2×2 block shows vector error magnitude and direction for Linear Interp., WeatherDCAE-14M, PixelAttn-VFI and WeatherBridge at the same query hour. Panel annotations give cosine-latitude-weighted vector RMSE (vRMSE) over the fixed 20−35°N, 80−105°W box. WeatherBridge has lower vRMSE than both learned comparators at four of five interior hours; WeatherDCAE-14M is lower at $\tau = 5$ h.

2 Twelve-hour per-channel errors

Supplementary Figs. S2 and S3 resolve the twelve-hour interval-length ablation by field. Model colours, markers and line styles match the six-hour comparison; values should be compared within a field because the physical scales differ.

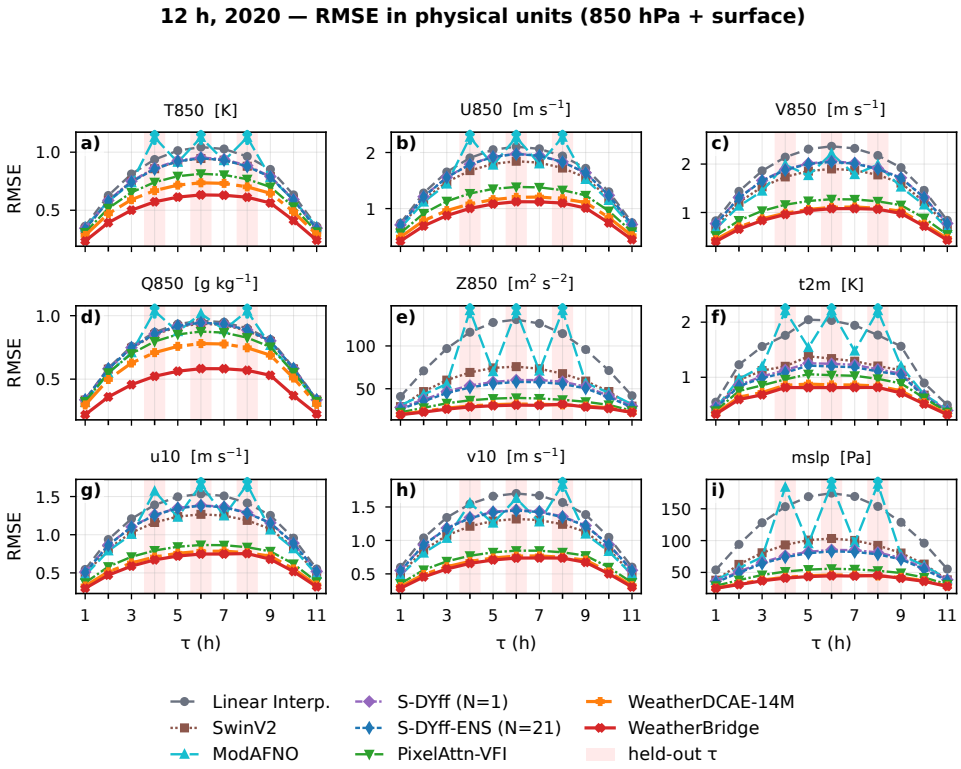


Fig. S2 Twelve-hour interval-length ablation for the 850-hPa and surface fields in physical units. Shading marks hours excluded from training; upward arrows mark clipped ModAFNO values beyond the displayed range.

Supplementary Fig. S4 resolves four representative same-trajectory results by forecast-lead bin; the aggregate and variable-group scores are reported in Supplementary Table S2 and summarised in the main text.

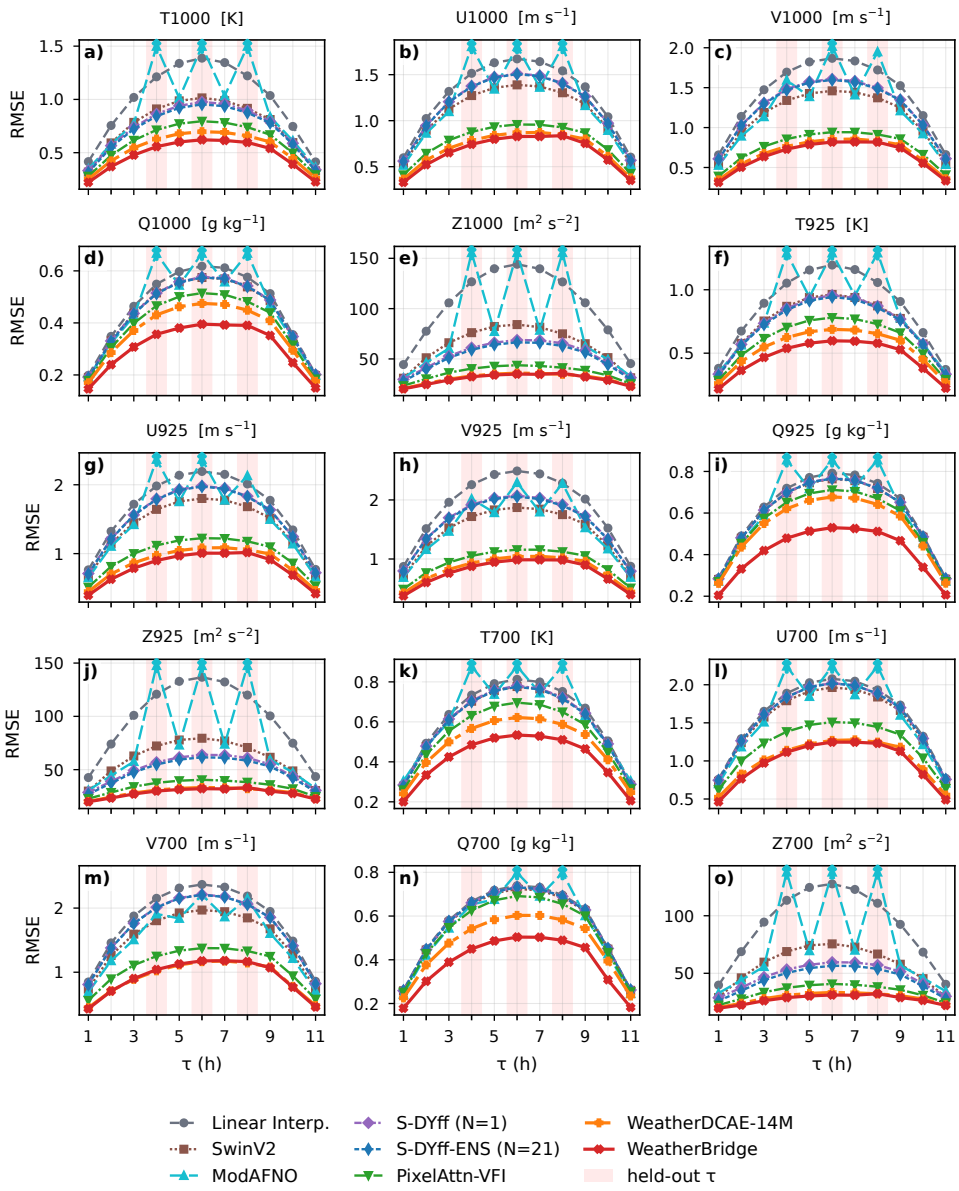


Fig. S3 Twelve-hour interval-length ablation for the remaining pressure-level fields in physical units. Shading marks hours excluded from training; upward arrows mark clipped ModAFNO values beyond the displayed range.

Table S1 Matched-v2 12 h interpolation on 2020 ERA5. Entries are mean ± standard deviation of the 24-channel macro RMSE across the indicated query hours. The standard deviation describes hour-to-hour variation, not sampling uncertainty over dates.

Model	Params (M)	Seen	Held out	All hours
WeatherBridge	14.3	**0.099 ± 0.033**	**0.133 ± 0.008**	**0.108 ± 0.032**
WeatherDCAE-14M	14.4	0.107 ± 0.034	0.141 ± 0.008	0.117 ± 0.033
PixelAttn-VFI	14.7	0.122 ± 0.038	0.160 ± 0.007	0.133 ± 0.037
SwinV2	7.9	0.168 ± 0.058	0.226 ± 0.011	0.184 ± 0.056
S-DYff	85.5	0.178 ± 0.061	0.240 ± 0.012	0.195 ± 0.059
ModAFNO	151.7	0.162 ± 0.056	0.369 ± 0.018	0.218 ± 0.108
Linear Interp.	–	0.202 ± 0.074	0.278 ± 0.015	0.223 ± 0.072

Table S2 Densifying a numerical forecast trajectory: both anchors and the target are states of one IFS HRES run, separated by twelve hours, and the model reconstructs the archived six-hour state between them. Sixteen 2021 initialisations, 39 windows each, 24 prognostic fields; scores are latitude-weighted normalised RMSE and per-group columns give the reduction relative to linear interpolation. All models are twelve-hour checkpoints evaluated on the identical index. Internal architecture variants appear in the separate ablation study.

Method	RMSE	vs. linear	T	U	V	Q	Z	surface
Linear interpolation	0.2897	—	—	—	—	—	—	—
PixelAttn-VFI	0.2033	+29.8%	+10.4	+30.2	+41.1	+5.7	−4.2	+29.6
WeatherDCAE-14M	0.1765	+39.1%	+20.1	+39.3	+49.4	+13.5	+15.5	+38.3
WeatherBridge	**0.1664**	**+42.6%**	+25.1	+42.5	+50.0	+32.2	+20.7	+40.4

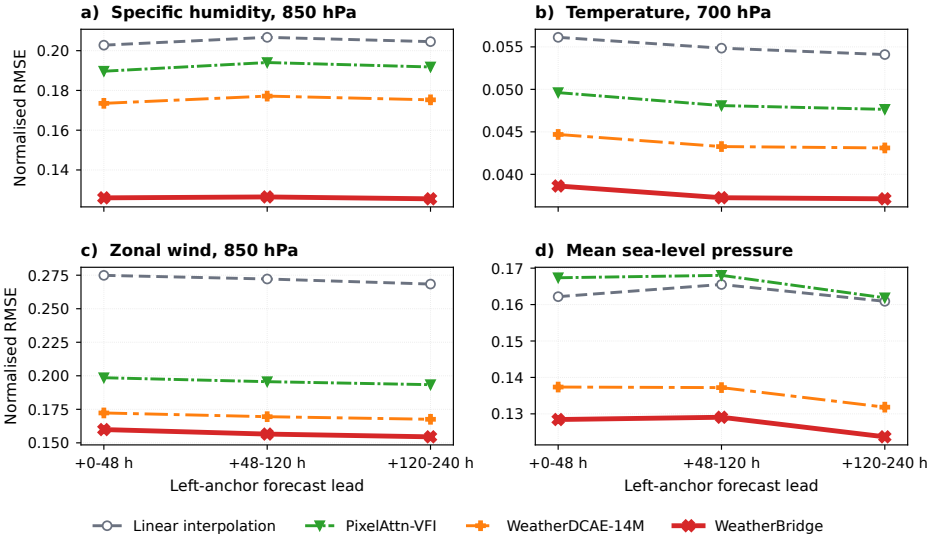


Fig. S4 Lead-resolved normalised RMSE for four representative fields in the same-trajectory HRES experiment. The x axis bins the forecast lead of the left anchor; all curves use the same 16 initialisations and archived targets. The panels were selected after evaluation as one field from each of four physical families with a large WeatherBridge margin over the strongest plotted competitor, so they are illustrative rather than a separate confirmatory test. The main text retains the aggregate 24-field result.

3 Calendar-year transfer audit

Supplementary Figs. S5–S6 use the same window construction, normalisation and query-hour split as the 2020 comparison. No learned model was trained on 2021, although these results were inspected during development and are not a post-selection test.

Table S3 Six-hour calendar transfer on the matched full-year 2021 index (7,295 window–hour samples). Seen hours are $\tau \in \{1, 3, 5\}$ and held-out hours are $\tau \in \{2, 4\}$. Entries are mean ± standard deviation of 24-channel macro scores across the indicated query hours; this is descriptive hour-to-hour variation, not sampling uncertainty over dates. Lower RMSE and higher ACC are better. † marks legacy evaluator records that share the frozen index but predate corrected-grid and dataset-identity provenance. They are kept for context and do not meet the current provenance standard.

Model	Seen RMSE	Held-out RMSE	Avg. RMSE	ACC
WeatherBridge	$\mathbf{0.055 \pm 0.019}$	$\mathbf{0.074 \pm 0.003}$	$\mathbf{0.062 \pm 0.017}$	$\mathbf{0.995 \pm 0.002}$
WeatherDCAE-14M	0.063 ± 0.021	0.081 ± 0.004	0.070 ± 0.018	0.993 ± 0.003
PixelAttn-VFI†	0.069 ± 0.022	0.091 ± 0.003	0.078 ± 0.019	0.992 ± 0.004
SwinV2†	0.076 ± 0.025	0.095 ± 0.003	0.083 ± 0.021	0.992 ± 0.004
S-DYff†	0.095 ± 0.033	0.119 ± 0.001	0.105 ± 0.026	0.990 ± 0.005
ModAFNO†	0.096 ± 0.035	0.121 ± 0.002	0.106 ± 0.028	0.988 ± 0.005
Linear Interp.	0.115 ± 0.040	0.144 ± 0.001	0.126 ± 0.033	0.986 ± 0.007

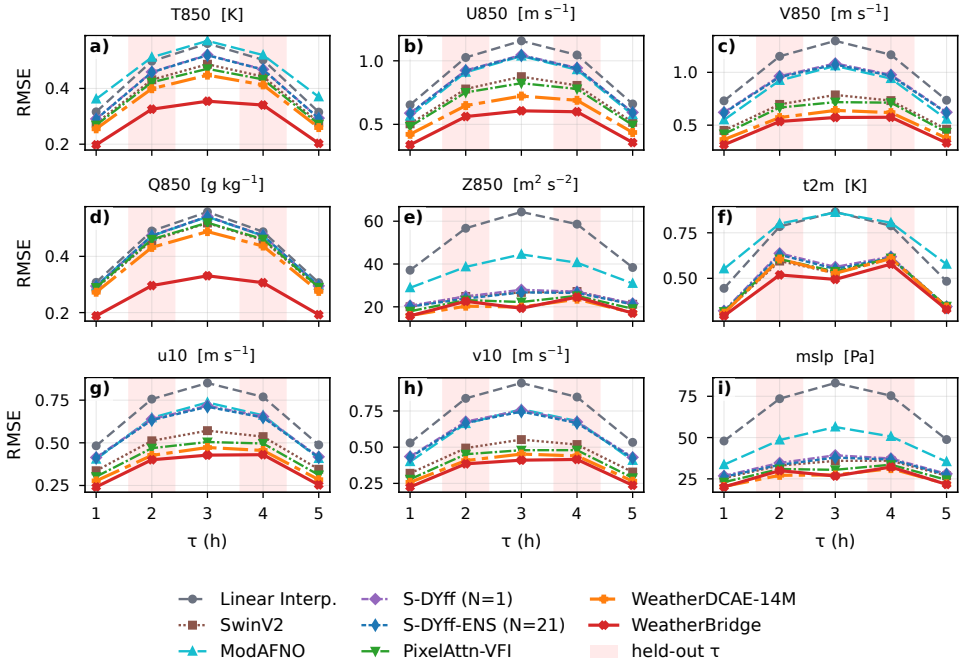


Fig. S5 Six-hour per-channel RMSE in physical units in 2021 for representative pressure-level and surface fields. Shading marks query hours excluded from training. WeatherBridge is red.

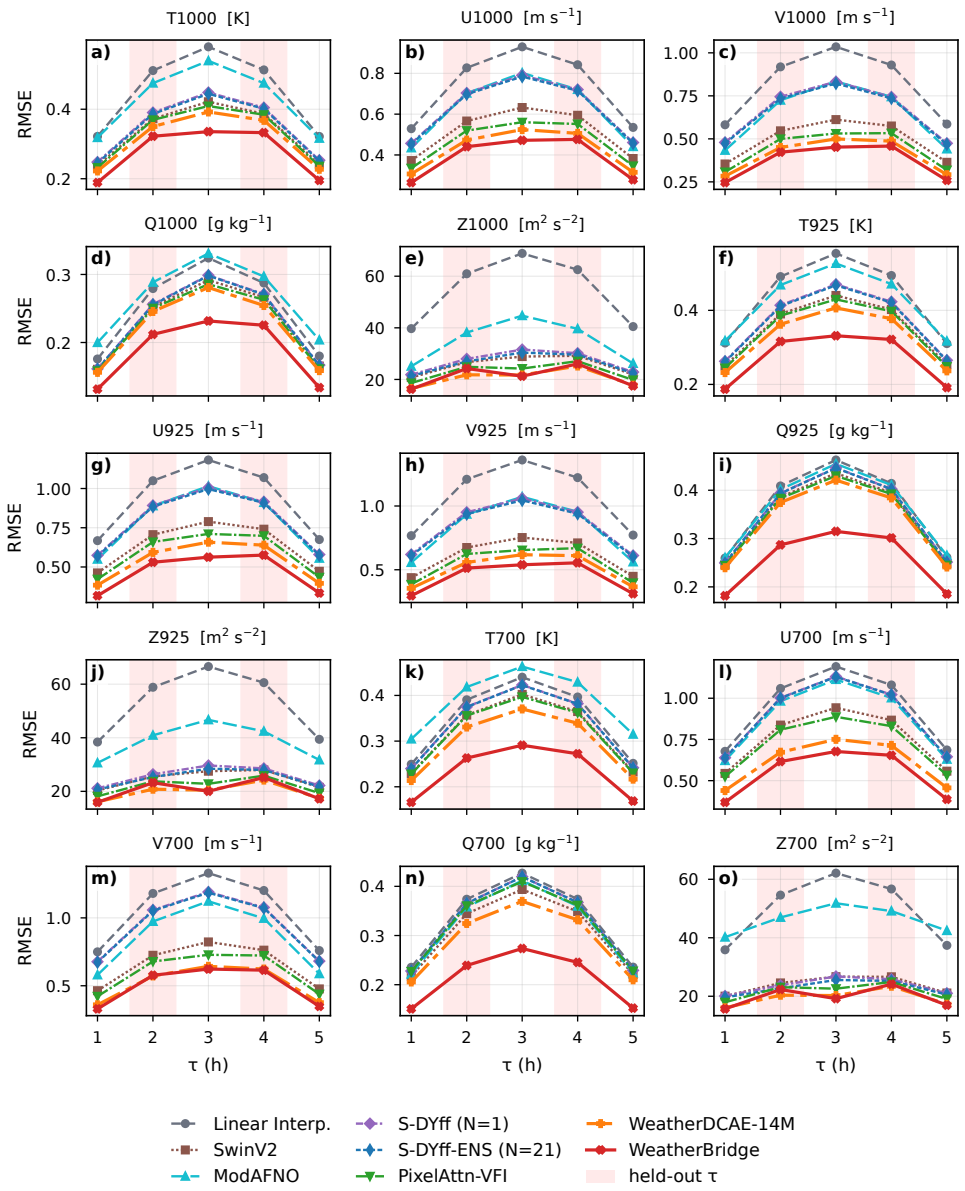


Fig. S6 Six-hour RMSE in physical units in 2021 for the remaining pressure-level channels. Shading marks query hours excluded from training.

4 Full-query supervision control

The principal experiment withholds query hours 2 and 4 during optimisation. Performance at those hours measures transfer. A counterpart trained on all five interior hours estimates the benefit of direct supervision. Both counterparts use the same 2014–2019 ERA5 pool, six-hour anchor construction, 24 fields and full-year 2020 and 2021 evaluation windows. Supplementary Table S4 records the checkpoint-level differences that determine which comparisons support a causal interpretation.

For WeatherBridge, seed, optimiser-step budget, effective batch size, precision and loss weights are matched; only the sampled query-hour set changes. Full-query training reduces mean normalised RMSE at hours 2 and 4 by 8.77% in 2020 and 8.78% in 2021 (Supplementary Fig. S7). At the trained hours 1, 3 and 5, it changes RMSE by -0.35% and -0.34%, respectively. Thus the extra supervision primarily improves the two hours omitted from sparse-query training. The full-query WeatherBridge remains the best model in absolute RMSE: 0.06015 in 2020 and 0.05991 in 2021 over all five hours, compared with 0.06466 and 0.06436 for WeatherDCAE-14M.

The WeatherBridge improvement at hours 2 and 4 is present in all 24 fields in both years. The largest reductions occur for geopotential (18.0–18.8%), MSLP (17.3–17.5%) and 2-m temperature (15.1–16.1%); wind fields improve by 6.5–10.5%, pressure-level temperature by 3.5–9.3% and specific humidity by 2.8–7.6%. The paired bootstrap defined in Methods gives a 95% interval of $[0.00650, 0.00655]$ for sparse-minus-full nRMSE in 2020 and $[0.00648, 0.00653]$ in 2021. No draw is non-positive, so the corresponding Monte Carlo tail probability is reported at its $1/5001$ resolution rather than as zero.

The response depends on the architecture (Supplementary Fig. S8). SwinV2 gains about 2.1% at hours 2 and 4, whereas S-DYff shows no gain. WeatherDCAE-14M and PixelAttn-VFI show larger gains, but their full-query checkpoints receive 21,888 updates rather than 13,136; those differences cannot be assigned to query-hour supervision alone. WeatherBridge provides the matched comparison. ModAFNO is omitted because its full-query run did not produce a complete checkpoint. Linear interpolation has no training condition and therefore no sparse-versus-full self-comparison. The S-DYff comparison uses single stochastic reconstructions from each checkpoint, with stochastic depth active at inference. It does not estimate an ensemble-mean effect; the inference protocol and its difference from original DYffusion are described in Methods.

Table S4 Provenance of the six-hour sparse-query and full-query checkpoints. The sparse-query runs draw $\tau \in \{1, 3, 5\}$; the full-query runs draw $\tau \in \{1, 2, 3, 4, 5\}$. Both variants use the same 2014–2019 ERA5 pool and are evaluated on identical 2020 and 2021 windows. Only the WeatherBridge pair matches both optimiser budget and verified seed.

Architecture	**Updates, sparse/full**	**Seed evidence**	**Use in this control**
WeatherBridge	13,136 / 13,136	Verified and matched	Primary matched self-comparison
WeatherDCAE-14M	13,136 / 21,888	Same seed	Diagnostic; extra optimisation confounds the query set
PixelAttn-VFI	13,136 / 21,888	Sparse seed not checkpoint-verified	Diagnostic; update budget and seed evidence differ
SwinV2	17,520 / 17,520	Sparse seed not checkpoint-verified	Diagnostic despite matched update counts
S-DYff	35,032 / 35,032	Sparse seed not checkpoint-verified	Diagnostic despite matched update counts

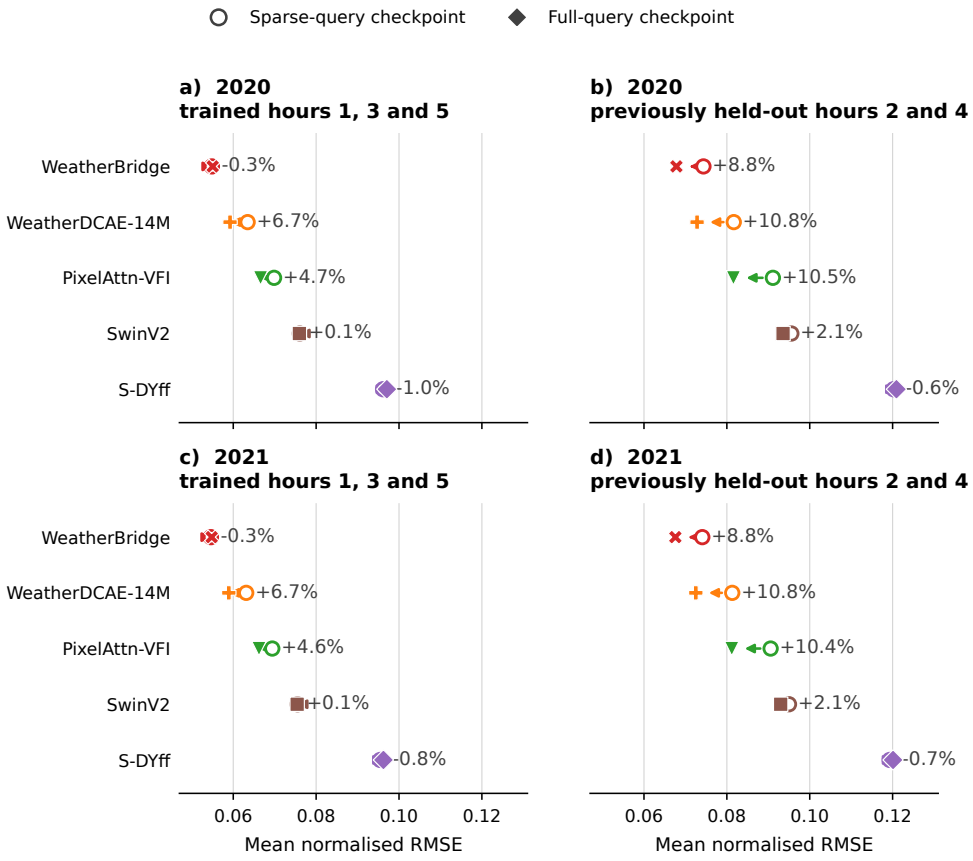


Fig. S7 Sparse-query versus full-query training on six-hour ERA5 intervals. Open circles denote checkpoints trained at hours 1, 3 and 5; filled symbols denote checkpoints trained at all five interior hours. Arrows point from the sparse-query score to the full-query score, and labels give the relative RMSE reduction. Lower normalised RMSE is better. Only WeatherBridge matches both the verified seed and update budget. The other architectures show diagnostic full-query sensitivity under the provenance constraints in Supplementary Table S4.

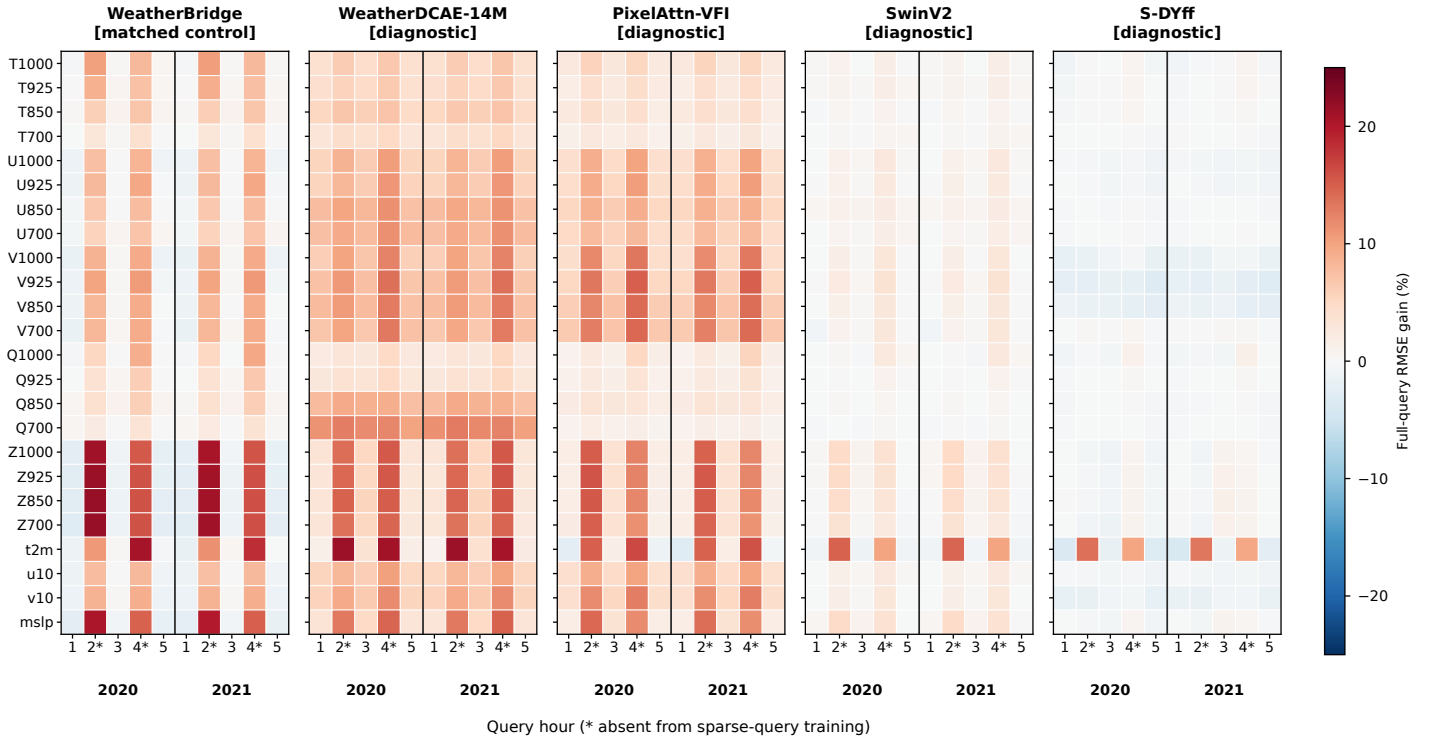


Fig. S8 Field-by-hour change from sparse-query to full-query training. Cells show $100(\mathrm{RMSE}_{\mathrm{sparse}} - \mathrm{RMSE}_{\mathrm{full}})/\mathrm{RMSE}_{\mathrm{sparse}}$; red is an improvement and blue a regression. Stars mark hours 2 and 4, which are absent from sparse-query training. Each architecture contains the 2020 and 2021 results on the same colour scale. WeatherBridge is the matched control; the remaining columns are diagnostic because their seed or update-budget provenance is incomplete or unequal.

5 Assimilation-window stratification

The full-year six-hour comparison contains four anchor starts. Pairs beginning at 00 and 12 UTC remain within one ERA5 4D-Var window, whereas those beginning at 06 and 18 UTC cross the 09 or 21 UTC boundary. The ordering is unchanged in all eight year–start strata, although WeatherBridge's advantage over WeatherDCAE-14M is smaller for boundary-crossing pairs. This audit limits the assimilation-window confound for RMSE. The spectral sample remains restricted to 00/12 UTC, and the 2022 sample to 00 UTC.

2020

Table S5 Six-hour RMSE stratified by anchor start time. Starts at 00 and 12 UTC keep both anchors in one ERA5 4D-Var window; starts at 06 and 18 UTC cross the 09 or 21 UTC boundary. Scores are mean channel-wise normalised RMSE on the paired full-year index.

Start UTC	WB	DCAE	Linear	WB vs DCAE
00 (within)	0.0608	0.0696	0.1274	+12.6%
06 (crosses 09 UTC)	0.0680	0.0752	0.1328	+9.5%
12 (within)	0.0616	0.0706	0.1290	+12.7%
18 (crosses 21 UTC)	0.0672	0.0747	0.1314	+10.0%
Within-window pooled	0.0612	0.0701	0.1282	+12.7%
Boundary-crossing pooled	0.0676	0.0749	0.1321	+9.7%

2021

Start UTC	WB	DCAE	Linear	WB vs DCAE
00 (within)	0.0605	0.0693	0.1269	+12.6%
06 (crosses 09 UTC)	0.0678	0.0748	0.1323	+9.4%
12 (within)	0.0613	0.0702	0.1283	+12.6%
18 (crosses 21 UTC)	0.0670	0.0744	0.1310	+9.9%
Within-window pooled	0.0609	0.0697	0.1276	+12.6%
Boundary-crossing pooled	0.0674	0.0746	0.1317	+9.6%

6 Conservative robustness sensitivity

The internal sensitivity defined in Methods checks fieldwise regressions against WeatherDCAE-14M alongside aggregate accuracy, parameter count, latency and HRES transfer. It was defined during development; it is reported separately from the primary held-out query-hour comparison.

This rule retains WeatherDCAE-14M because WeatherBridge has local regressions despite lower aggregate and held-out RMSE. Supplementary Table S6 gives the

temporal-curvature outcomes. Supplementary Table S7 gives representative vector-SHT values; Supplementary Data 1 contains the complete families.

Table S6 Selector diagnostics for WeatherBridge against WeatherDCAE-14M. Lower is better. Curvature intervals use paired seven-day blocks ($n = 53$). In the physical diagnostics (Supplementary Data 1), lower-tropospheric moisture bias is the sole Holm-significant regression in every row.

Interval, year	WeatherBridge curvature	WeatherDCAE curvature	Δ [95% CI]
6 h, 2020	0.08570	0.08477	+0.00094 [+0.00076, +0.00111]
6 h, 2021	0.08533	0.08427	+0.00106 [+0.00090, +0.00122]
12 h, 2020	0.09048	0.08784	+0.00264 [+0.00257, +0.00271]
12 h, 2021	0.08980	0.08719	+0.00262 [+0.00254, +0.00268]

Table S7 Vector-SHT signed cospectrum for 850-hPa wind in 2020. One is ideal; higher is better. Query hours include one held-out and one trained time per interval. Intervals are paired over calendar-week blocks. Supplementary Data 1 contains the complete multiplicity-corrected vector family.

Interval	τ	Component	WeatherBridge	WeatherDCAE	Δ [95% CI]
6 h	2	Spheroidal	0.973	0.966	+0.007 [+0.006, +0.007]
6 h	2	Toroidal	0.881	0.853	+0.027 [+0.027, +0.028]
6 h	3	Spheroidal	0.969	0.958	+0.011 [+0.010, +0.011]
6 h	3	Toroidal	0.862	0.821	+0.041 [+0.041, +0.042]
12 h	5	Spheroidal	0.852	0.828	+0.024 [+0.023, +0.025]
12 h	5	Toroidal	0.546	0.523	+0.022 [+0.021, +0.024]
12 h	8	Spheroidal	0.860	0.833	+0.027 [+0.026, +0.028]
12 h	8	Toroidal	0.588	0.561	+0.027 [+0.026, +0.028]

7 Architecture ablation study

Table S8 Transport-architecture ablation. RMSE is the 24-field full-year mean; lower is better. Internal variants are confined to this audit table; the main comparison uses WeatherBridge.

Variant	Architectural difference	Params (M)	6 h 2020	6 h 2021	12 h 2020
WeatherBridge	Transport trunk + spectral supervision	14.3	0.0626	0.0624	0.1081
Detail	+ transported anchor-detail bypass	14.3	0.0625	0.0622	0.1090
Refine	Latent transformer; shared heads; no mass head	13.7	0.0624	0.0622	0.1096

The retrained Refine variant, defined in Methods, lowered aggregate six-hour ERA5 RMSE, but two field–hour regressions remained in the inspected 2021 transfer year: temperature (T2m) regressions at $\tau = 1$ and $\tau = 5$, with one-sided Holm-adjusted $p = 0.024$ each. It also lost all three paired IFS HRES seeds, and the conservative robustness sensitivity rejects it. Supplementary Data 1 retains its configuration, checkpoints and paired statistics.

The same sensitivity also rejects Detail, the anchor-detail bypass variant defined in Methods.

Refine and Detail lower six-hour full-year ERA5 RMSE by less than one per cent relative to WeatherBridge, but transfer less reliably across data sources. Refine loses all three paired IFS HRES seeds at both horizons; Detail loses its only completed seed pair at both horizons, leaving its cross-source evidence incomplete. WeatherBridge has the lowest twelve-hour ERA5 RMSE of the three and wins all three twelve-hour HRES seed pairs. We retain it as the representative transport formulation. The stricter multi-metric rule retains WeatherDCAE-14M as the conservative reference.

The retrained branches above change several mechanisms at once. To isolate components used by the retained checkpoint, we also apply same-checkpoint inference lesions. Disabling the acceleration term increases aggregate RMSE by 2.16%, while forcing the warped-anchor gate to zero increases it by 60.54%. Disabling the hydrostatic head increases aggregate RMSE by only 0.04%, as it changes five of 24 channels, but all 25 affected geopotential/MSLP field–hour cells worsen after Holm correction. Averaged over the five query hours, hydrostatic-balance MSE increases by 1.06% (Supplementary Table S9, panel b). The transport-off arm retains the learned residual corrections and isolates warped-anchor transport without reducing the model to linear interpolation. No arm is retrained.

Table S9 Same-checkpoint inference lesions on the full-year 2020 six-hour index. No arm is retrained. Positive Δ and percentage values denote degradation relative to intact WeatherBridge. All three aggregate lesion tests reach the Monte Carlo resolution limit (1/5001) under paired seven-day-block permutation. Panel b resolves the hydrostatic lesion over geopotential Z at four pressure levels and mean sea-level pressure (MSLP); RMSE and percentage change are computed from averages across the five query hours, followed by the minimum–maximum hourly change in brackets.

a. Aggregate lesions

Inference arm	RMSE	Δ vs intact	Change (%)
Intact WeatherBridge	0.06263	+0.00000	+0.00
Acceleration disabled	0.06398	+0.00135	+2.16
Warped-anchor transport disabled	0.10055	+0.03792	+60.54
Hydrostatic head disabled	0.06265	+0.00003	+0.04

b. Hydrostatic-head target fields

Field	Intact RMSE	Head-off RMSE	Change (%)
Z_{1000}	0.02326	0.02342	+0.69 [+0.24, +1.06]
Z_{925}	0.01846	0.01857	+0.59 [+0.21, +0.87]
Z_{850}	0.01441	0.01450	+0.66 [+0.27, +0.97]
Z_{700}	0.00938	0.00945	+0.77 [+0.25, +1.15]
MSLP	0.02360	0.02378	+0.74 [+0.31, +1.13]

8 Anchor-order diagnostic

Neither principal architecture is equivariant to exchanging the two anchors. For WeatherBridge, the exchange RMSE is 0.287 in 2020 and 0.285 in 2021, respectively 3.49 and 3.44 times its forward reconstruction error. The corresponding WeatherDCAE-14M ratios are 3.01 and 2.98. "Bidirectional" refers to the two predicted transport directions. The complete interpolation map is not reversible because the input stack and learned path assign distinct roles to the left and right anchors.

Table S10 Anchor-exchange diagnostic on the six-hour task. Exchange RMSE compares $f(x_0, x_T, t)$ with $f(x_T, x_0, 1-t)$; forward RMSE compares the first prediction with ERA5. Lower is better. The 48-endpoint, 240-window sample per year is diagnostic and was not used for selection.

Model	Year	Exchange RMSE	Forward RMSE	Ratio
WeatherBridge	2020	0.287	0.082	3.49
WeatherBridge	2021	0.285	0.083	3.44
WeatherDCAE-14M	2020	0.278	0.092	3.01
WeatherDCAE-14M	2021	0.276	0.093	2.98

9 Twelve-hour spectral amplitude

Supplementary Fig. S9 shows spectral amplitude. The scalar and vector-SHT tables below report the signed cospectrum and vector-field tests needed to assess coefficient alignment.

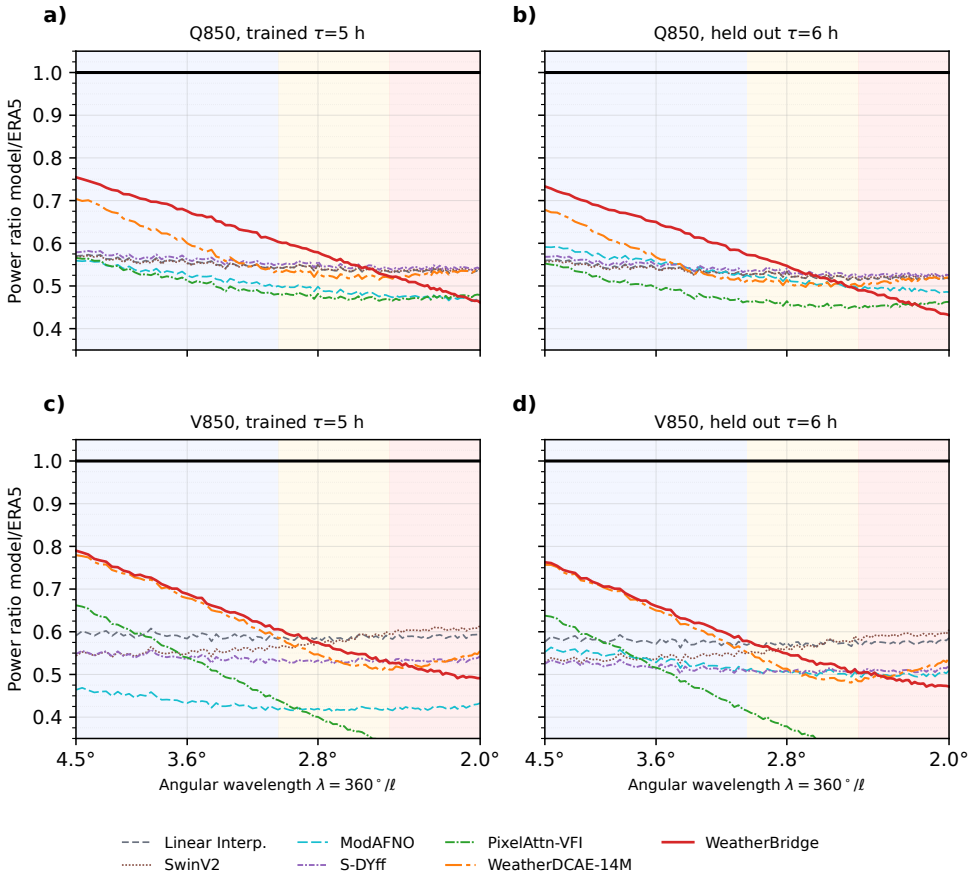


Fig. S9 Twelve-hour spectral power retention. Rows show Q850 and V850; columns compare a trained query hour ($\tau = 5$) with a held-out hour ($\tau = 6$). Curves show the model-to-ERA5 power ratio over $\ell \in [80, 180]$; one is ideal. WeatherBridge has the lowest mean absolute log-ratio error for Q850 in both columns and is shown in solid red.

Table S11 Per-channel spherical-harmonic high-frequency preservation ratio $R_{\mathrm{HF}} = \sum_{\ell \geq 80} E_{\mathrm{pred}}(\ell) / \sum_{\ell \geq 80} E_{\mathrm{GT}}(\ell)$ for the 6 h leaderboard, averaged over $\tau \in \{1, 2, 3, 4, 5\}$ across all 24 prognostic channels (5 PL variables × 4 levels = 20 + 4 surface). $R_{\mathrm{HF}} = 1$ = perfect spectral fidelity; values < 1 indicate HF energy under-prediction. Per-row **best** (closest to 1.0) and worst (farthest from 1.0) highlighted. Block means aggregate over the 4 PL levels per variable; the surface block mean uses the 4 surface fields.

Channel	WeatherBridge	WeatherDCAE-14M	PixelAttn-VFI	SwinV2	ModAFNO	S-DYff	Linear Interp.
T1000	**0.884**	0.879	0.877	0.871	0.874	0.878	0.851
T925	**0.857**	0.843	0.826	0.825	0.818	0.823	0.796
T850	**0.811**	0.791	0.768	0.776	0.773	0.777	0.751
T700	0.771	**0.773**	0.765	0.765	0.772	0.768	0.753
mean(T)*	0.831	0.821	0.809	0.809	0.809	0.812	0.788
U1000	**0.900**	0.893	0.855	0.841	0.804	0.818	0.771
U925	**0.889**	0.889	0.851	0.836	0.792	0.801	0.756
U850	**0.872**	0.867	0.804	0.806	0.772	0.780	0.742
U700	**0.823**	0.821	0.729	0.753	0.727	0.735	0.704
mean(U)*	0.871	0.867	0.810	0.809	0.774	0.784	0.743
V1000	**0.911**	0.899	0.863	0.842	0.781	0.809	0.749
V925	**0.902**	0.890	0.849	0.830	0.752	0.783	0.730
V850	**0.891**	0.885	0.817	0.817	0.757	0.761	0.718
V700	**0.852**	0.846	0.760	0.757	0.708	0.709	0.672
mean(V)*	0.889	0.880	0.822	0.812	0.750	0.766	0.717
Q1000	0.854	0.851	**0.857**	0.847	0.854	0.857	0.837
Q925	**0.798**	0.762	0.759	0.751	0.755	0.762	0.738
Q850	**0.827**	0.762	0.743	0.736	0.740	0.754	0.715
Q700	**0.764**	0.743	0.731	0.736	0.740	0.752	0.716
mean(Q)*	0.811	0.780	0.773	0.768	0.772	0.781	0.751
Z1000	0.928	**0.938**	0.920	0.930	0.924	0.935	0.901
Z925	0.918	**0.928**	0.913	0.921	0.914	0.922	0.884
Z850	0.899	**0.911**	0.895	0.903	0.893	0.898	0.857
Z700	0.844	**0.868**	0.831	0.857	0.865	0.826	0.759
mean(Z)*	0.897	0.911	0.890	0.903	0.899	0.895	0.850
t2m	0.926	0.932	0.948	0.932	0.938	**0.953**	0.920
u10	**0.902**	0.896	0.858	0.841	0.803	0.818	0.771
v10	**0.911**	0.901	0.865	0.842	0.777	0.807	0.746
mslp	0.925	**0.934**	0.917	0.925	0.921	0.931	0.899
mean(surface)	0.916	0.916	0.897	0.885	0.860	0.877	0.834
mean (24ch)	0.869	0.863	0.833	0.831	0.811	0.819	0.781

Table S12 Per-channel spherical-harmonic high-frequency preservation ratio $R_{\mathrm{HF}} = \sum_{\ell \geq 80} E_{\mathrm{pred}}(\ell) / \sum_{\ell \geq 80} E_{\mathrm{GT}}(\ell)$ for the 12 h leaderboard, averaged over $\tau \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$ across all 24 prognostic channels (5 PL variables × 4 levels = 20 + 4 surface). $R_{\mathrm{HF}} = 1$ = perfect spectral fidelity; values < 1 indicate HF energy under-prediction. Per-row **best** (closest to 1.0) and worst (farthest from 1.0) highlighted. Block means aggregate over the 4 PL levels per variable; the surface block mean uses the 4 surface fields.

Channel	WeatherBridge	WeatherDCAE-14M	PixelAttn-VFI	SwinV2	ModAFNO	S-DYff	Linear Interp.
T1000	0.794	**0.818**	0.792	0.804	0.802	0.810	0.802
T925	0.771	**0.781**	0.743	0.754	0.755	0.762	0.752
T850	0.731	**0.743**	0.704	0.726	0.733	0.732	0.723
T700	0.712	**0.753**	0.744	0.737	0.740	0.744	0.733
mean(T)*	0.752	0.774	0.746	0.755	0.757	0.762	0.752
U1000	0.753	**0.795**	0.710	0.721	0.700	0.731	0.728
U925	0.773	**0.812**	0.734	0.708	0.691	0.721	0.715
U850	0.744	**0.775**	0.650	0.703	0.684	0.704	0.703
U700	0.667	**0.694**	0.546	0.681	0.658	0.681	0.679
mean(U)*	0.735	0.769	0.660	0.703	0.683	0.709	0.706
V1000	0.775	**0.794**	0.720	0.693	0.666	0.716	0.718
V925	0.790	**0.811**	0.718	0.675	0.642	0.692	0.709
V850	0.742	**0.750**	0.625	0.678	0.630	0.664	0.695
V700	0.649	0.652	0.506	0.644	0.579	**0.679**	0.669
mean(V)*	0.739	0.752	0.642	0.672	0.629	0.688	0.698
Q1000	0.767	**0.798**	0.777	0.780	0.777	0.785	0.780
Q925	**0.719**	0.714	0.672	0.698	0.675	0.702	0.699
Q850	**0.721**	0.695	0.631	0.664	0.663	0.673	0.665
Q700	**0.673**	0.668	0.641	0.662	0.660	0.668	0.661
mean(Q)*	0.720	0.719	0.680	0.701	0.694	0.707	0.701
Z1000	0.874	0.909	0.875	0.898	0.903	**0.933**	0.890
Z925	0.872	0.909	0.875	0.892	0.908	**0.928**	0.879
Z850	0.856	0.899	0.860	0.873	**0.935**	0.914	0.847
Z700	0.801	0.871	0.829	0.813	1.242	**0.888**	0.732
mean(Z)*	0.851	0.897	0.860	0.869	0.997	0.916	0.837
t2m	0.846	0.896	0.882	0.893	0.893	**0.899**	0.890
u10	0.755	**0.800**	0.711	0.719	0.697	0.729	0.727
v10	0.777	**0.797**	0.721	0.689	0.658	0.711	0.714
mslp	0.870	0.902	0.869	0.891	0.902	**0.924**	0.884
mean(surface)	0.812	0.849	0.796	0.798	0.787	0.816	0.804
mean (24ch)	0.768	0.793	0.731	0.750	0.758	0.766	0.750

10 Diagnostic physical and regional summaries

The balance ratios use 180 uniformly spaced windows, while the diurnal errors aggregate four regions, 12 months and three surface fields. These reduced-sample diagnostics are descriptive; the full-year physical suite supplies the paired robustness analysis.

Table S13 Diagnostic-balance ratios relative to ERA5 over 180 uniformly spaced 2020 anchor windows and all interior hours. One is the reference value; bold marks the closest ratio within each horizon and diagnostic.

Horizon	Model	Ageo 850	Ageo 700	Hydrostatic
6 h	WeatherBridge	0.949	0.940	0.992
	WeatherDCAE-14M	**0.962**	**0.962**	**0.994**
	Linear Interp.	0.936	0.903	0.993
12 h	WeatherBridge	0.924	0.920	0.985
	WeatherDCAE-14M	**0.958**	**0.981**	**0.992**
	Linear Interp.	0.926	0.878	**0.992**

Table S14 Regional diurnal-cycle errors over four regions, 12 months and three surface fields (144 strata). Amplitude error is $|A_{\mathrm{recon}}/A_{\mathrm{ERA5}} - 1|$; peak error is the circular local-time difference. Lower is better; bold marks the lowest value per horizon.

Horizon	Model	Amplitude error	Peak error (h)
6 h	WeatherBridge	**0.043**	**0.507**
	WeatherDCAE-14M	0.047	0.674
	Linear Interp.	0.170	1.569
12 h	WeatherBridge	**0.068**	**0.562**
	WeatherDCAE-14M	0.070	0.660
	Linear Interp.	0.508	3.597

11 Training and architecture inventory

Tables S15 and S16 record the training budgets and architecture settings of the reported checkpoints.

Table S15 Training protocol for the reported checkpoints. Models grouped within a row share the training years, epoch budget and optimiser-step count. Every run uses the same 24 fields and peak learning rate 10^{-4}.

Setting	Years	Epochs	Updates
WeatherBridge, WeatherDCAE-14M, PixelAttn-VFI (6 h)	2014–2019	8	13,136
SwinV2 (6 h)	2014–2019	8	17,520
ModAFNO, S-DYff (6 h)	2014–2019	8	35,032
WeatherBridge, WeatherDCAE-14M, PixelAttn-VFI (12 h)	2017–2019	10	10,930
SwinV2, S-DYff (12 h)	2017–2019	10	5,470
ModAFNO (12 h)	2017–2019	10	21,870

Table S16 Per-variant architecture settings. Parameter counts come from the reported checkpoint state dictionaries; *n-modes* gives the spherical-FNO truncation (latitude×longitude).

Variant	Params	Hyperparameters
WeatherBridge	14.3 M	hidden 72; three-level circular-padded U-Net; second-order transport, blend, pyramid residual, SFNO and mass-field heads; replicated-latitude high-pass boundary; FFT-magnitude term on the advected-channel mask
WeatherDCAE-14M	14.4 M	block-out 64/128/256, layers 3/3/3, latent 256; longitude-periodic, scalar-style pole-reflected convolutions
SwinV2	7.9 M	embed 256, depth 8, heads 8, window 5×9, patch 4
ModAFNO (6 h)	36.9 M	embed 256, depth 8, blocks 8, MLP ratio 2
ModAFNO (12 h)	151.7 M	embed 512, depth 12, blocks 8, MLP ratio 2
S-DYff (6 h)	101.3 M	embed 128, depth 4, n-modes 16×32, K=5 refine
S-DYff (12 h)	85.5 M	embed 96, depth 6, n-modes 16×32, K=5 refine
PixelAttn-VFI	14.7 M	hidden 72; three levels; shared anchor encoder; bottleneck self-attention and AdaLN-Zero; mean-endpoint decoder skips

12 Forecast-anchor spectral audit

The four-cell audit below accompanies the adapted IFS HRES result in the main text. It separates small-scale amplitude from coefficient alignment: improved coherence and signed cospectrum do not imply uniformly improved spectral energy or shape.

Table S17 Independent 2022 forecast-anchor spectral differences over $\ell = 80$–180. Values are the frozen coefficient adapter minus Linear Interp.; negative is better for energy and shape error, whereas positive is better for coherence and signed cospectrum. † denotes Holm-adjusted $p < 0.05$ over the two fields within each metric and query hour.

τ	Field	Δ energy error ↓	Δ shape error ↓	Δ coherence ↑	Δ signed cospectrum ↑
2	Q850	+0.0009†	+0.0023†	+0.0064†	+0.0069†
2	U850	+0.0044†	−0.0008†	+0.0045†	+0.0041†
3	Q850	−0.0006	+0.0093†	+0.0112†	+0.0122†
3	U850	+0.0091†	−0.0001	+0.0108†	+0.0099†

13 Coefficient adapter by field

Supplementary Fig. S10 uses 16 IFS HRES initialisations chosen before the 2022 evaluation. Pairs of states separated by six hours are drawn from the same forecast trajectory and used as the two anchors. Linear interpolation, WeatherDCAE-14M, WeatherBridge and its frozen coefficient adapter reconstruct the five intervening hourly states; each prediction is compared with ERA5 at the same valid time. Here, τ denotes the offset within an HRES anchor pair rather than forecast lead time. These scores include both interpolation error and the mismatch between HRES and ERA5, so they are not conventional HRES forecast-skill curves.

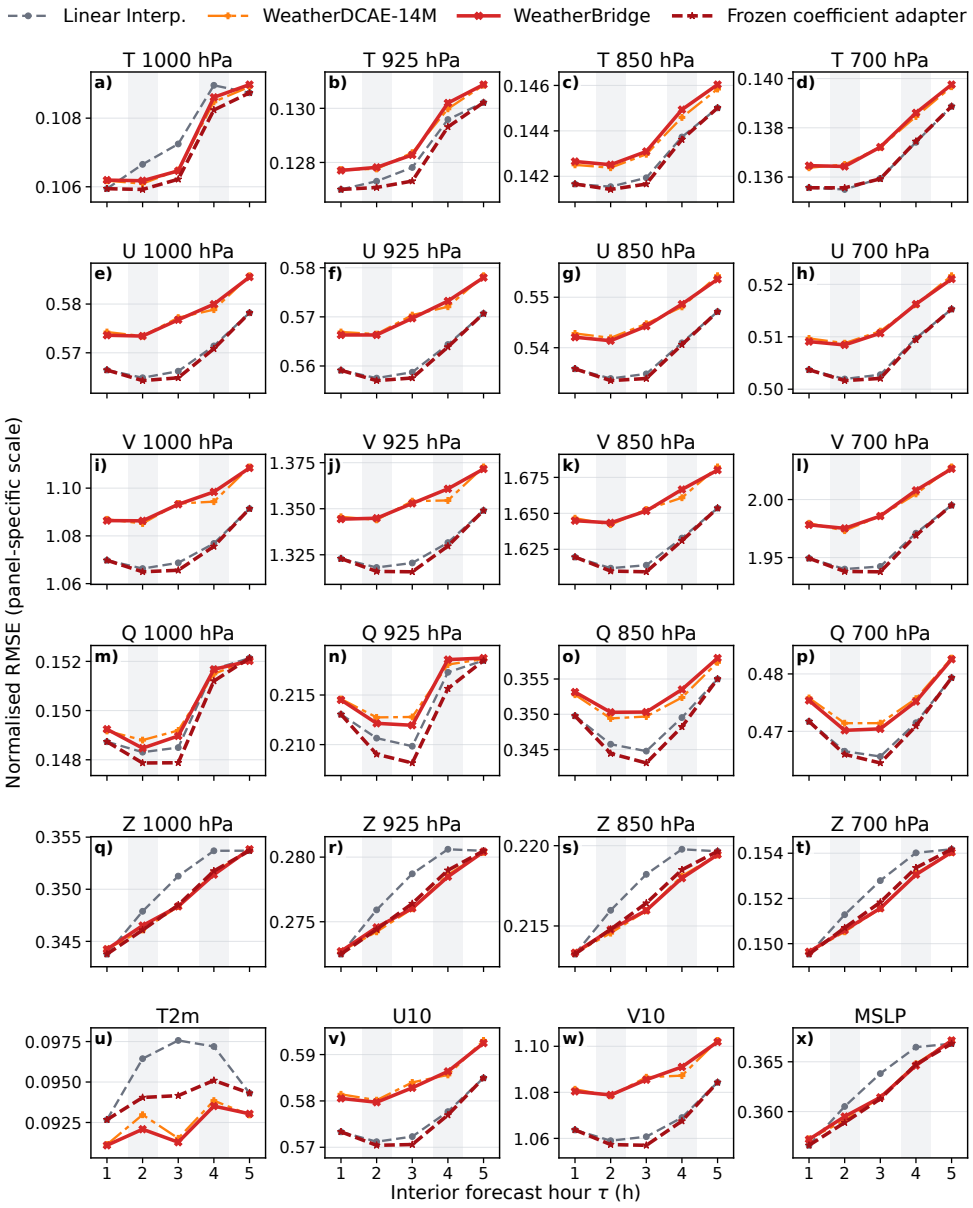


Fig. S10 IFS HRES anchor interpolation against ERA5 for 16 independent 2022 initialisations. The 24 panels follow the five hours inside a six-hour anchor interval; τ is interior offset, not forecast lead. Curves show Linear Interp., WeatherDCAE-14M, raw WeatherBridge and the frozen coefficient adapter. Grey bands mark the held-out hours $\tau \in \{2, 4\}$; ranges are panel-specific. Only the adapter-versus-Linear comparison is confirmatory. The endpoint guard ties $\tau \in \{1, 5\}$ exactly; T700 at $\tau = 2$ and 4 contains the only regressions.

14 Reference architecture schematics

Supplementary Fig. S11 summarises the comparator architectures defined in Methods.

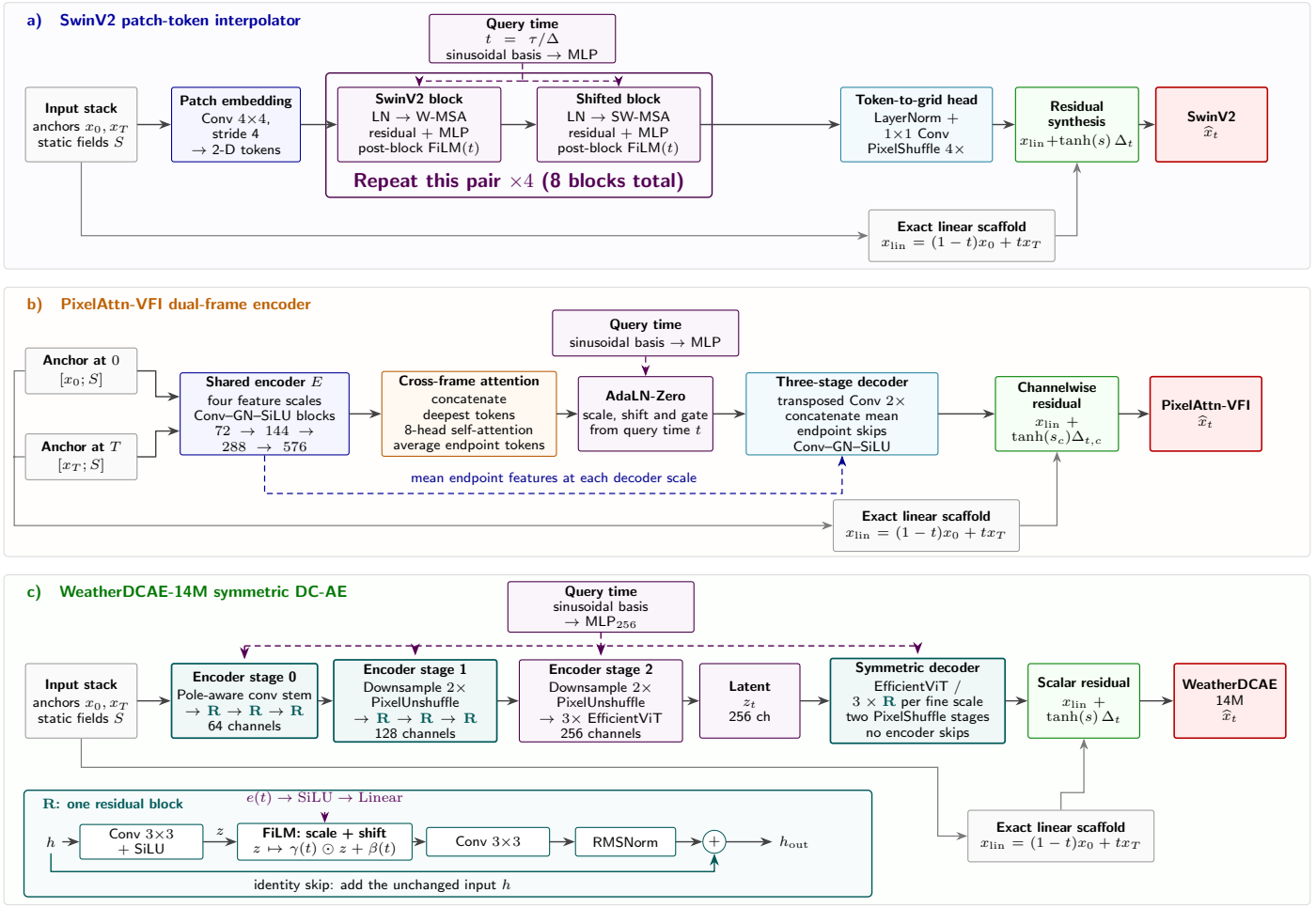


Fig. S11 Reference architectures used in the principal comparison. **a)** SwinV2 alternates windowed and shifted-window attention on patch tokens, with time-conditioned FiLM after each block. **b)** PixelAttn-VFI shares the endpoint encoder, mixes deepest tokens through attention, and conditions the bottleneck on time. Mean endpoint features supply decoder skips. **c)** WeatherDCAE-14M has no cross-scale encoder–decoder skips. Read stages top to bottom: Conv–PixelUnshuffle halves height and width, then three blocks run sequentially. Channel counts denote feature width. Teal R blocks occur at 64 and 128 channels in both encoder and decoder; the deepest stage uses EfficientViT multiscale linear attention. The inset expands one R: Conv–SiLU, FiLM, Conv and RMSNorm act sequentially, then add the unchanged input h through the identity skip. FiLM applies channelwise scale and shift derived from time embedding $e(t)$. Convolutions in R use pole-aware boundaries. Purple dashed arrows carry time conditioning; blue carries endpoint skips. Each model adds a scaled correction to linear interpolation.

15 S-DYff ensemble averaging

S-DYff-ENS averages 21 stochastic reconstructions from the S-DYff checkpoint. The plotted S-DYff curve retains the original single reconstruction used in the tables. For a paired estimate of the averaging effect, we also recomputed $N = 1, 4, 16, 21$ using nested prefixes of the same random draws. Relative to this paired $N = 1$ control, $N = 21$ reduces aggregate normalised RMSE by 0.69–0.89% across the six year–interval combinations and by 0.67–0.87% at held-out hours. The paired control is a new draw, so these percentages need not equal the difference from the archived S-DYff curve. Seven-day block-bootstrap intervals describe variation across evaluation dates conditional on the sampling seed, not variation across sampling seeds.

All 24 fields are shown for full-year 2020 and 2021 and the fixed 48-anchor 2022 sample. The six-hour 2020, twelve-hour 2020 and six-hour 2021 panels appear earlier in the main text and Supplementary Information; the remaining three comparisons appear in Supplementary Figs. S12, S13 and S14. Ensemble results cover RMSE and ACC only. The spectral, HRES and full-query supervision experiments retain their original inference settings.

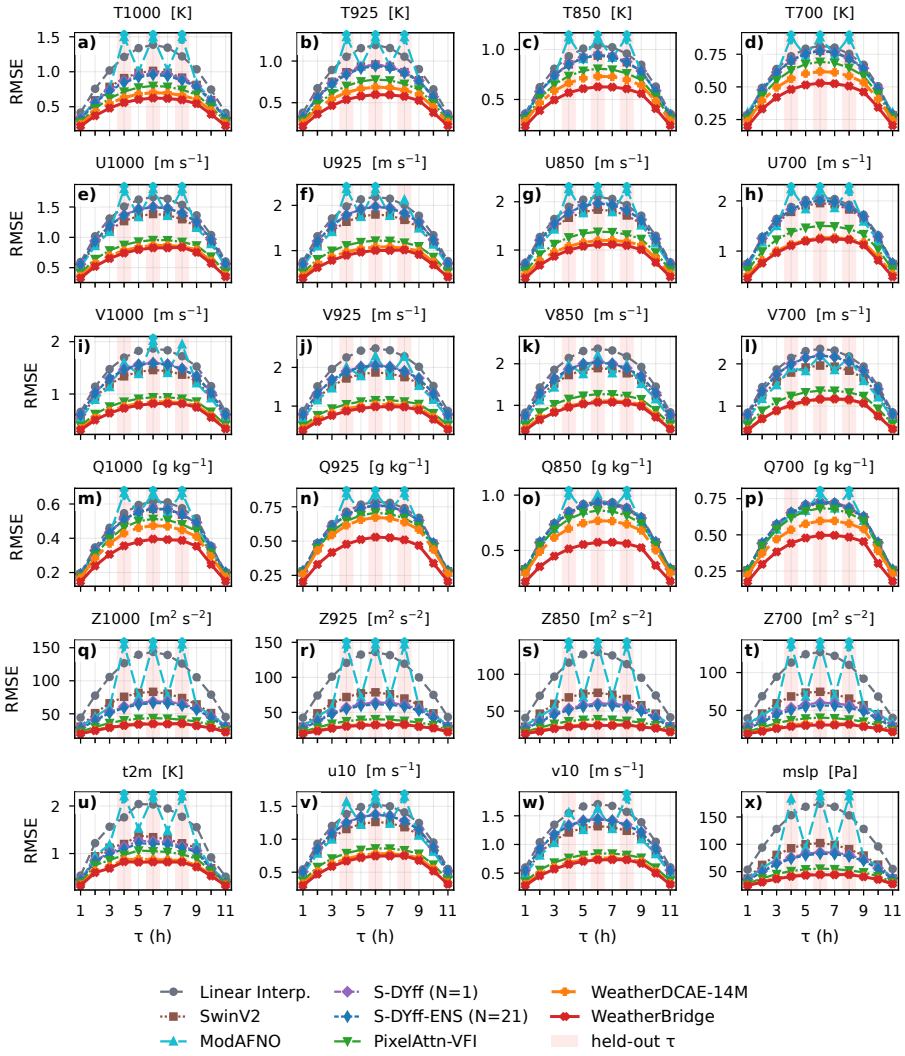


Fig. S12 Twelve-hour physical-unit RMSE for all 24 fields over full-year 2021. S-DYff ($N = 1$) and S-DYff-ENS ($N = 21$) use the same checkpoint. Shading marks held-out query hours; upward arrows mark clipped ModAFNO values.

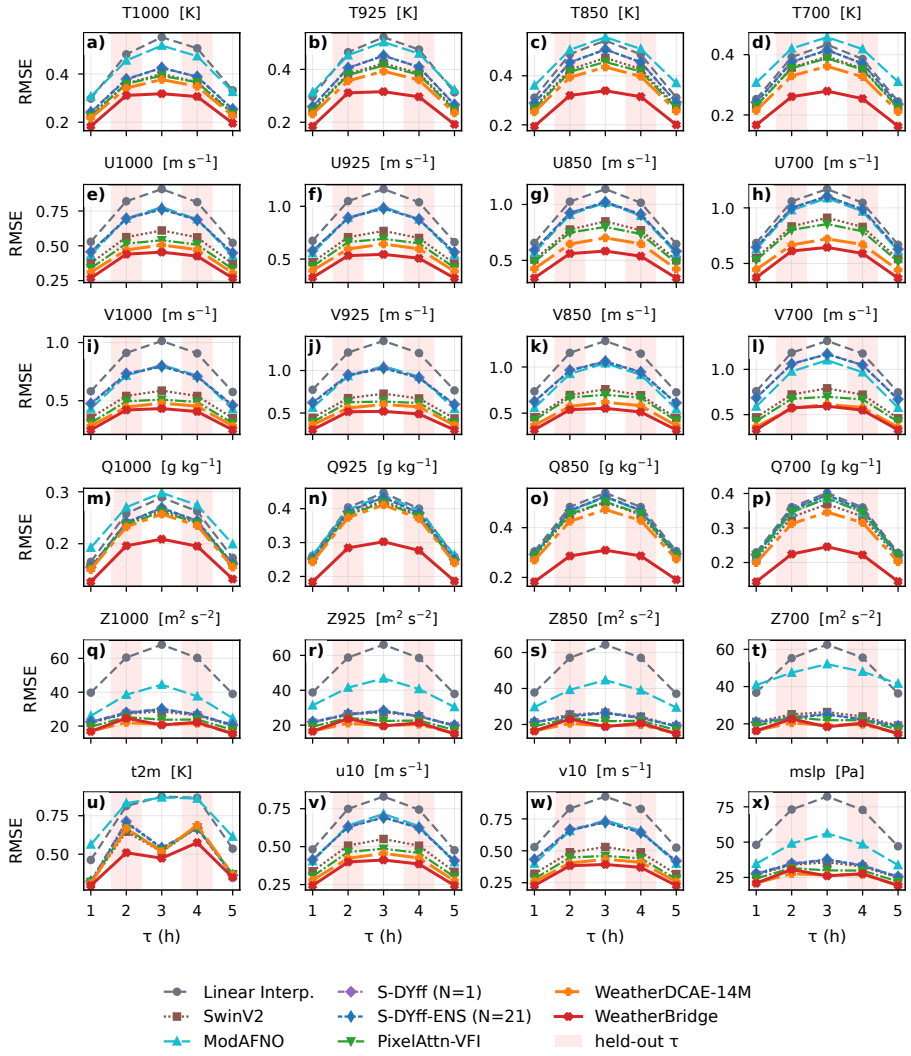


Fig. S13 Six-hour physical-unit RMSE for all 24 fields on the fixed 48-anchor 2022 sample. S-DYff ($N = 1$) is the archived single reconstruction; S-DYff-ENS ($N = 21$) averages outputs from the same checkpoint. Shading marks held-out query hours.

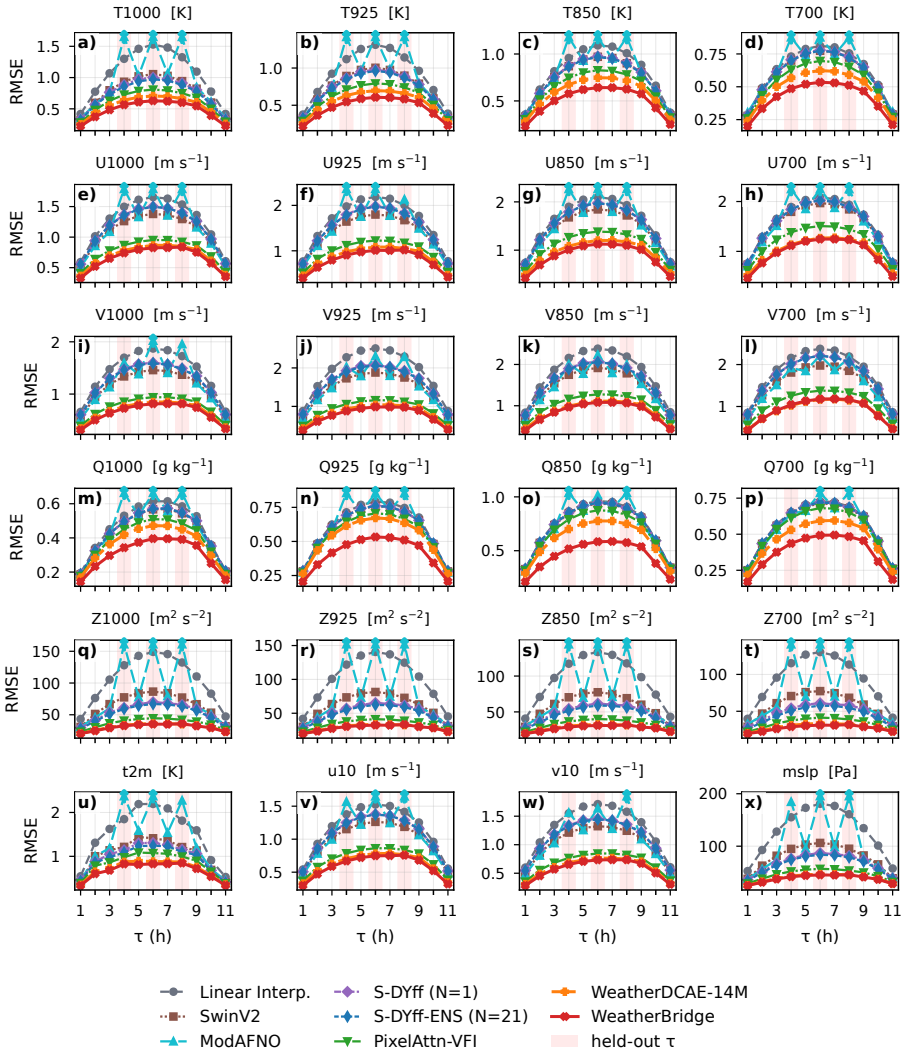


Fig. S14 Twelve-hour physical-unit RMSE for all 24 fields on the fixed 48-anchor 2022 sample. Model labels follow Supplementary Fig. S13. Shading marks held-out query hours; upward arrows mark clipped ModAFNO values.