

Supplementary Information: Transfer of neural
ensemble temperature post-processing to
unobserved stations: random and spatially blocked
evaluation

Yi Zou^{1*}

^{1*}Independent researcher, Xi'an, China.

Corresponding author(s). E-mail(s): yizou505@gmail.com;

Online Resource 1 for *Meteorology and Atmospheric Physics*. These materials retain the original reference analyses; every ensemble size and exploratory status remains labelled. No new experiment was performed in preparing this resource.

Resource overview.

Tables [S1](#), [S2](#), [S3](#), [S4](#), [S5](#), [S6](#), [S7](#), [S8](#), [S9](#), [S10](#), [S11](#) retain calibration, original comparisons, exploratory subsets, pilot results and sensitivity analyses. Figures [S1](#), [S2](#), [S3](#), [S4](#), [S5](#), [S6](#), [S7](#), [S8](#), [S9](#) retain the original three-seed comparison and individual station-diagnostic panels.

047 **S1 Additional data, method and result details**

048

049 *German predictors.*

050

051 The ensemble variables are those of Table 1 of [Rasp and Lerch \(2018\)](#): 2 m temperature,
052 convective available potential energy, surface pressure, total cloud cover, sensible and
053 latent heat flux, 10 m u and v wind, 2 m dew point, short- and long-wave radiation flux,
054 soil moisture, u and v wind at 500 and 850 hPa, geopotential at 500 hPa and specific
055 humidity at 850 hPa. [Rasp and Lerch \(2018\)](#) describe each variable by its ensemble
056 mean and standard deviation; the published file stores mean and variance, which we
057 use as provided. Soil moisture is dropped because of missing values, as in the authors'
058 replication code.
059
060
061
062
063

064

065 *EUPPBench predictors and download.*

066

067 Ensemble mean and standard deviation of t2m, u10, v10, tcc, sd, stl1, swvl1, tcwv,
068 cape, the 6-h window variables mx2t6, mn2t6, sshf6, slhf6, ssr6, str6 and 850 hPa
069 temperature; sine and cosine of day of year; a lead-time one-hot vector; and latitude,
070 longitude, station altitude, model orography, alt – orog and a one-hot CORINE level-1
071 land-use class. Models see only ensemble statistics, never individual members. Instead
072 of the full 18.3 GB archive we fetched only the required Zarr stores and checked all
073 45 346 files against the sizes and CRC32 checksums of the Zenodo listing. The stations
074 excluded by the 50% availability rule are two Belgian, one French and two German
075 stations.
076
077
078
079
080
081

082

083 *11-member sensitivity (EUPPBench, stage-specified for the Stage-3* 084 *methods).*

085

086 Computing the test-period ensemble statistics from members 0–10 only, matching
087 the reforecast ensemble size, made every post-processing method worse by 0.0074–
088 0.0086 °C on random and 0.0062–0.0100 °C on spatial folds (raw ensemble +0.022 °C).
089
090
091
092

The ranking on random folds is unchanged; on spatial folds NN-UNK and NN-Hybrid-Res[†] swap places 3 and 4 (51 members: 0.9832 vs. 0.9803; 11 members: 0.9894 vs. 0.9903 °C).

DRN and GNN-Geo details.

The DRN uses the features all ensemble predictors, lat, lon, alt, orog, $\sin(2\pi \text{doy}/365.25)$, day of year and month (plus the lead one-hot on EUPPBench), with early stopping as for the other networks. GNN-Geo uses the same node features; a batch of 256 time slices corresponds to 256 dates (German) or initialisation–lead pairs (EUPPBench); training runs for at most 100 epochs with patience 5. Isolated nodes use only the root term $W_{\text{root}}\mathbf{x}_i$. Keeping unobserved stations out of the training graph is stricter than in [Lakatos \(2026\)](#).

Table S1 Calibration at held-out stations (same 3-seed ensembles as Table S3): empirical coverage of the central 80% interval (nominal 0.80) and spread–skill ratio (RMS σ / RMSE of μ ; 1 = consistent).

Method	German				EUPPBench			
	random		spatial		random		spatial	
	cov80	SSR	cov80	SSR	cov80	SSR	cov80	SSR
Raw ensemble	0.395	0.43	0.395	0.43	0.512	0.58	0.512	0.58
EMOS	0.801	0.94	0.798	0.94	0.817	0.98	0.811	0.97
Boosted EMOS	0.816	1.00	0.817	1.00	0.844	1.07	0.838	1.19
NN-UNK	0.748	0.86	0.751	0.86	0.788	0.94	0.776	0.94
NN-kNN	0.753	0.87	0.755	0.88	0.785	0.94	0.778	0.94
NN-NoEmb	0.774	0.92	0.771	0.91	0.787	0.94	0.745	0.89
NN-Attr	0.770	0.91	0.749	0.87	0.786	0.93	0.729	0.76
NN-Hybrid	0.761	0.88	0.743	0.85	0.791	0.94	0.739	0.79
NN-Hybrid-kNN [†]	0.754	0.87	0.735	0.84	0.791	0.94	0.739	0.79
NN-Hybrid-Res [†]	0.761	0.89	0.745	0.85	0.793	0.95	0.781	0.95
DRN	0.810	0.98	0.808	0.98	0.848	1.06	0.803	0.85
GNN-Geo	0.835	1.05	0.828	1.04	0.843	1.06	0.820	1.01
SAMOS-MLP	0.777	0.93	0.769	0.91	0.796	0.96	0.761	0.87
SAMOS-Attn	0.734	0.83	0.749	0.86	0.777	0.92	0.772	0.83

Table S2 Stage-specified primary tests of NN-Hybrid (German and EUPPBench protocols). Random folds: 10-seed ensembles; spatial folds: 3-seed ensembles. DM test on daily means (Newey–West lag 5), BH over the 6 tests per dataset and fold type. “st.+”/“st.–”: fraction of stations whose own station-level DM test (BH across stations) is significant in favour of / against NN-Hybrid (stage-specified station-level analysis).

Data	Folds	vs B	Δ	rel. %	DM p	BH	st.+	st.–
German	random	EMOS	-0.1577	-15.62	$< 10^{-15}$	yes	0.85	0.03
German	random	Boosted EMOS	-0.0605	-6.64	$< 10^{-15}$	yes	0.66	0.06
German	random	NN-UNK	-0.0127	-1.47	2.2e-16	yes	0.40	0.24
German	random	NN-kNN	-0.0280	-3.19	0.00015	yes	0.43	0.33
German	random	NN-NoEmb	-0.0031	-0.37	0.00054	yes	0.32	0.19
German	random	NN-Attr	+0.0000	+0.00	1	no	0.26	0.22
German	spatial	EMOS	-0.1171	-11.56	$< 10^{-15}$	yes	0.73	0.07
German	spatial	Boosted EMOS	-0.0198	-2.16	0.00041	yes	0.48	0.14
German	spatial	NN-UNK	+0.0037	+0.42	0.1	no	0.34	0.29
German	spatial	NN-kNN	-0.0111	-1.22	0.13	no	0.40	0.31
German	spatial	NN-NoEmb	+0.0183	+2.08	6.7e-16	yes	0.25	0.33
German	spatial	NN-Attr	-0.0167	-1.83	8.9e-16	yes	0.32	0.24
EUPPBench	random	EMOS	-0.1889	-16.96	$< 10^{-15}$	yes	0.85	0.09
EUPPBench	random	Boosted EMOS	-0.0375	-3.89	$< 10^{-15}$	yes	0.74	0.17
EUPPBench	random	NN-UNK	+0.0129	+1.41	$< 10^{-15}$	yes	0.42	0.42
EUPPBench	random	NN-kNN	+0.0060	+0.65	0.0042	yes	0.40	0.44
EUPPBench	random	NN-NoEmb	-0.0135	-1.43	$< 10^{-15}$	yes	0.53	0.20
EUPPBench	random	NN-Attr	-0.0118	-1.25	$< 10^{-15}$	yes	0.41	0.37
EUPPBench	spatial	EMOS	-0.0264	-2.32	0.015	yes	0.56	0.30
EUPPBench	spatial	Boosted EMOS	+0.1361	+13.93	$< 10^{-15}$	yes	0.42	0.48
EUPPBench	spatial	NN-UNK	+0.1299	+13.21	$< 10^{-15}$	yes	0.28	0.56
EUPPBench	spatial	NN-kNN	+0.1329	+13.56	$< 10^{-15}$	yes	0.29	0.66
EUPPBench	spatial	NN-NoEmb	+0.0371	+3.45	$< 10^{-15}$	yes	0.37	0.51
EUPPBench	spatial	NN-Attr	-0.0429	-3.71	$< 10^{-15}$	yes	0.49	0.34

S1.1 Deviations from the protocols

1. *Boosted EMOS*. The first implementation boosted the raw temperature and had not converged after the specified 1000 iterations (validation CRPS still falling). Standardising the response inside the booster fixed this with unchanged settings. The problem was found on validation data, but we had also seen the 2016 score of the local boosted EMOS on seen stations (0.8675 before, 0.8020 after the fix; 0.80 in Rasp and Lerch 2018).

2. <i>Primary design.</i> The protocols named random folds as primary. Because the difference between fold types became the object of study after the results were known, this paper gives both fold types <i>equal weight</i> ; this is a post hoc change of emphasis.	185 186 187 188 189
3. <i>Compute environment.</i> The EUPPBench protocol planned GPU training. After a partial GPU run, all EUPPBench networks were retrained on one 8-core CPU machine so that no seed ensemble mixes devices; the 61 GPU runs were discarded.	190 191 192 193 194
4. <i>DRN and GNN-Geo.</i> Fitted once instead of with a rolling refit; GNN-Geo excludes held-out stations from the training graph.	195 196 197
5. <i>Station count.</i> The German protocol refers to “all 537 station ids”; the data file contains 535, all of which are used.	198 199 200 201
6. <i>Pilot evaluation.</i> An evaluator bug crashed the automatic Phase-1 check of the pilot, after which the launcher skipped Phase 2; the bug was fixed and the evaluation rerun on unchanged predictions (Section S5).	202 203 204 205 206
The diagnosis of out-of-range stations (Section 5.3 (main article)) was not specified in any protocol. Its definition was implemented after the German and EUPPBench results were known but before the Lakatos-style baselines and the pilot were specified and trained.	207 208 209 210 211 212 213 214
S1.2 Significance tests	215 216
For the original reference comparisons, we average the CRPS over all held-out stations (and leads) for each test day, giving 366 daily mean differences on the German data and 730 on EUPPBench. We apply the Diebold and Mariano (1995) test with a Newey and West (1987) variance estimate (lag 5) and a 95% moving-block bootstrap interval with 7-day blocks and 10 000 replicates (Künsch 1989). Averaging over stations first accounts for the spatial correlation of errors on a given day, but these tests treat the set of held-out stations as fixed: they do not capture the variability due to which stations and regions happen to be held out, nor seed variability (Section 6.3 (main article)). The	217 218 219 220 221 222 223 224 225 226 227 228 229 230

stage-specified German and EUPPBench protocols additionally specified station-level DM tests with BH correction across stations, which we report as fractions of stations. Within each pre-specified family we control the false discovery rate with the Benjamini and Hochberg (1995) procedure at $q = 0.05$. The families are (i) the six primary comparisons of NN-Hybrid (German and EUPPBench protocols), (ii) the seven pre-specified *descriptive* comparisons involving the DRN and GNN-Geo, corrected within each dataset and fold type, and (iii) the six primary pilot tests per dataset. All other tests, including those in the exploratory analysis, are uncorrected and descriptive.

Supplementary inference.

For each comparison, the primary supplementary interval resamples seeds within each method, rebuilds its ensemble, and then resamples whole stations (2000 replicates). Each replicate uses the ratio of total score differences to total rows, preserving the pooled-CRPS estimand. For within-design model differences, a separate interval resamples whole folds with seed resampling (2000 replicates); with seven unequal regions, this interval is coarse. It is not a paired random-spatial DiD interval; paired regional block intervals for the DiD were computed post hoc (Section S6.5). We also report fixed-ensemble station intervals (10 000 replicates) and descriptive station/fold sign and Wilcoxon tests. Bootstrap probabilities are fractions of resampled replicates, not posterior probabilities of superiority.

Supplementary daily-series DM tests use automatic Bartlett bandwidth selection (Newey and West 1994) and the small-sample adjustment of Harvey et al. (1997), with maximum lead 2 days (German) or 5 days (EUPPBench). Sensitivities use the Andrews (1991) bandwidth, fixed lags 5, 20 and 40, fixed- b inference (Kiefer and Vogelsang 2005), and moving blocks of 7, 14 and 30 days (10 000 replicates). BH correction is applied within each dataset/fold-type family to 17 specified pairs: ten networks

and the two new primary statistical comparators against Boosted EMOS, DRN/GNN-Geo/NN-UNK/SAMOS-MLP against NN-Hybrid-Res, and EMOS-RegIDW against EMOS. NN-Hybrid-kNN remains outside this family.

For $F \in \{\text{DRN}, \text{GNN-Geo}\}$ and $E = \text{Boosted EMOS}$, define

$$\text{DiD}_F = (\text{CRPS}_F - \text{CRPS}_E)_{\text{spatial}} - (\text{CRPS}_F - \text{CRPS}_E)_{\text{random}}. \quad (1)$$

Station resampling is paired across designs, while seeds are resampled independently. Positive DiD denotes a loss of relative skill under spatial blocking. The supplementary wording rules require both EUPPBench DiD intervals to exclude zero and positive point estimates for each method in at least four of six additional partitions to support this relative deterioration (C1). A ranking reversal requires the random and spatial station intervals separately to exclude zero with opposite signs (C2). The internal descriptive condition C3 additionally required both German DiD intervals to include zero or lie below it. This records a pattern of significance and signs across datasets; it is not a test of their effect difference and cannot establish a density contrast. Pair-wise superiority requires both a seed-nested station interval excluding zero and BH rejection by the primary DM test (C4); if the within-design region interval includes zero, the claim is restricted to the evaluated, record-weighted station network. No ordering is stated when the bootstrap superiority fraction is between 0.1 and 0.9 (C5). Any added statistical baseline with lower spatial CRPS than every network must be reported (C6). These rules govern the supplementary conclusions; the earlier three-seed analyses are retained as labelled reference results.

Table S3 Mean CRPS ($^{\circ}\text{C}$) at held-out stations, pooled over the 7 station-held-out folds (every station is held out exactly once). **All networks: 3-seed ensembles** (seeds 0–2; mean of μ and σ); single-seed mean $\pm$ sd in parentheses. Bold: lowest CRPS in each original three-seed column (raw ensemble excluded); this is descriptive, not a superiority claim. Ten-seed spatial scores and inference are in Table 2 (main article) and Section 5.2 (main article). German: test year 2016, lead 48 h; EUPPBench: 2017–2018, leads 24–120 h. † Declared exploratory in the protocols.

Method	German		EUPPBench	
	random	spatial	random	spatial
Raw ensemble	1.1519	1.1519	1.2113	1.2113
EMOS	1.0097	1.0130	1.1139	1.1395
Boosted EMOS	0.9125	0.9158	0.9625	0.9770
NN-UNK	0.8706 (0.882 $\pm$ 0.001)	0.8922 (0.905 $\pm$ 0.007)	0.9205 (0.935 $\pm$ 0.007)	0.9832 (1.000 $\pm$ 0.019)
NN-kNN	0.8849 (0.900 $\pm$ 0.007)	0.9070 (0.924 $\pm$ 0.007)	0.9276 (0.945 $\pm$ 0.009)	0.9801 (1.001 $\pm$ 0.022)
NN-NoEmb	0.8580 (0.866 $\pm$ 0.002)	0.8777 (0.886 $\pm$ 0.001)	0.9427 (0.955 $\pm$ 0.006)	1.0759 (1.099 $\pm$ 0.025)
NN-Attr	0.8563 (0.867 $\pm$ 0.004)	0.9126 (0.927 $\pm$ 0.008)	0.9392 (0.974 $\pm$ 0.020)	1.1560 (1.213 $\pm$ 0.092)
NN-Hybrid	0.8565 (0.867 $\pm$ 0.005)	0.8960 (0.913 $\pm$ 0.018)	0.9309 (0.953 $\pm$ 0.010)	1.1131 (1.147 $\pm$ 0.033)
NN-Hybrid-kNN †	0.8660 (0.876 $\pm$ 0.005)	0.9067 (0.923 $\pm$ 0.020)	0.9313 (0.953 $\pm$ 0.010)	1.1119 (1.145 $\pm$ 0.034)
NN-Hybrid-Res †	0.8531 (0.862 $\pm$ 0.001)	0.8977 (0.912 $\pm$ 0.007)	0.9134 (0.930 $\pm$ 0.001)	0.9803 (1.009 $\pm$ 0.008)
DRN	0.8645 (0.872 $\pm$ 0.002)	0.8795 (0.887 $\pm$ 0.003)	0.9170 (0.923 $\pm$ 0.001)	1.0787 (1.090 $\pm$ 0.089)
GNN-Geo	0.8840 (0.897 $\pm$ 0.004)	0.9048 (0.917 $\pm$ 0.005)	0.9468 (0.967 $\pm$ 0.005)	1.0423 (1.075 $\pm$ 0.053)
SAMOS-MLP	0.8544 (0.867 $\pm$ 0.001)	0.8784 (0.891 $\pm$ 0.004)	0.9198 (0.938 $\pm$ 0.007)	1.0023 (1.026 $\pm$ 0.021)
SAMOS-Attn	0.8738 (0.895 $\pm$ 0.004)	0.8865 (0.906 $\pm$ 0.004)	0.9364 (0.956 $\pm$ 0.020)	1.0218 (1.051 $\pm$ 0.109)

S2 Original three-seed comparisons and seen-to-new gaps

S2.1 Seen-to-new-station performance gap

Networks with a free per-station embedding (NN-UNK, NN-kNN, NN-Hybrid, NN-Hybrid-kNN † , NN-Hybrid-Res †) are the best models on seen stations (0.783–0.788 $^{\circ}\text{C}$ German, 0.822–0.826 $^{\circ}\text{C}$ EUPPBench; Table S4). Their pooled CRPS is 8.5–11.7% higher at new than seen German stations and 10.9–12.1% higher on EUPPBench. The two networks without a free embedding behave differently on the two datasets. On German data their gap is small (NN-NoEmb 0.9%, NN-Attr 1.9%) but they are worse on seen stations (0.848 and 0.836 $^{\circ}\text{C}$); on EUPPBench their gap (12.4% and 13.1%) is as large as that of the embedding networks. The global statistical models have much smaller gaps: 0.1% and 0.9% for Boosted EMOS, 0.05% and 0.3% for EMOS. Within a fold, seen and new scores use different stations and model–station combinations. The pooled gap is descriptive; it does not isolate the effect of removing one fixed

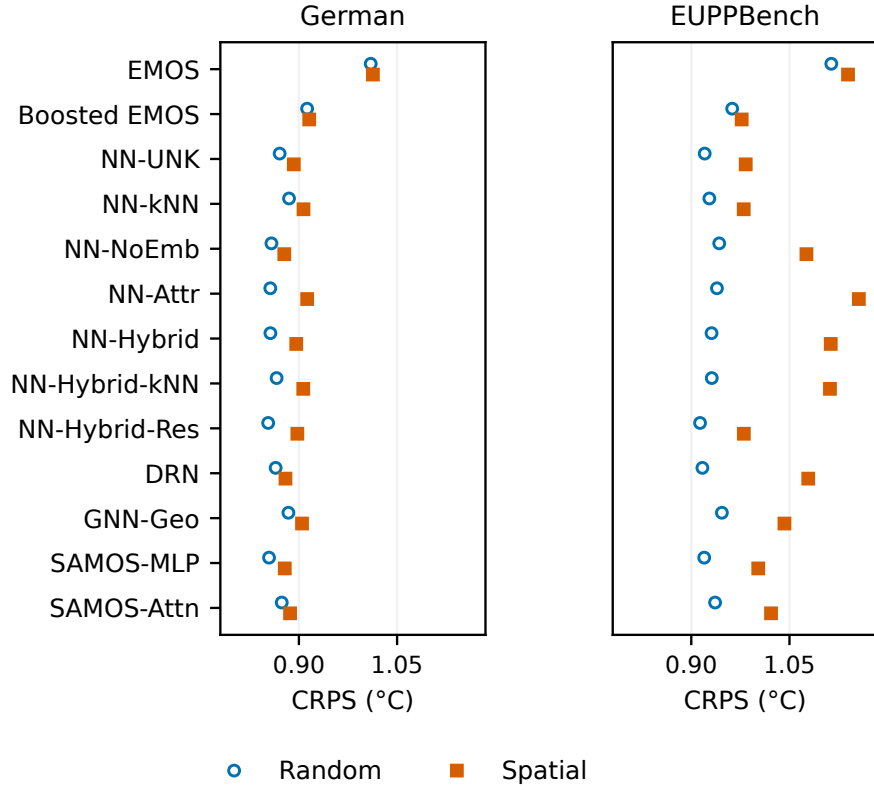


Fig. S1 Original three-seed ensemble CRPS at held-out stations for random (blue open circles) and spatially blocked (orange filled squares) folds. The raw ensemble is omitted from the plotting range. Scores for all methods and single-seed variability for the networks are retained in Table S3. NN-Hybrid-kNN and NN-Hybrid-Res are exploratory.

station’s training history. For orientation, a boosted EMOS fitted separately for each seen station, not available at a new station, scores 0.8020°C (German) and 0.8523°C (EUPPBench).

S2.2 Random folds

On random folds all networks have lower pooled CRPS than Boosted EMOS, by 1.6–6.5% (Table S3). On the German data the five best methods – NN-Hybrid-Res[†] (0.8531°C), SAMOS-MLP (0.8544), NN-Attr (0.8563), NN-Hybrid (0.8565) and NN-NoEmb (0.8580) – lie within 0.005°C of each other, followed by the DRN (0.8645).

Table S4 Cost of cold start, random folds: CRPS ($^{\circ}\text{C}$) on the fold’s own training stations (“seen”, mean over the 7 folds, test period) versus held-out stations (“new”). Gap = new/seen -1 .
Networks: 10-seed ensembles (seeds 0–9), the design of the German and EUPPBench protocols; the “new” values therefore differ slightly from the 3-seed values of Table S3. Seen and new scores are not computed on the same stations (each station is “seen” in 6 of 7 folds).

Method	German			EUPPBench		
	seen	new	gap	seen	new	gap
EMOS	1.009	1.010	+0.0%	1.110	1.114	+0.3%
Boosted EMOS	0.911	0.913	+0.1%	0.954	0.963	+0.9%
NN-UNK	0.787	0.865	+9.9%	0.822	0.912	+11.0%
NN-kNN	0.788	0.880	+11.7%	0.822	0.919	+11.8%
NN-NoEmb	0.848	0.855	+0.9%	0.835	0.939	+12.4%
NN-Attr	0.836	0.852	+1.9%	0.829	0.937	+13.1%
NN-Hybrid	0.785	0.852	+8.5%	0.826	0.925	+12.0%
NN-Hybrid-kNN [†]	0.785	0.862	+9.8%	0.826	0.925	+12.1%
NN-Hybrid-Res [†]	0.783	0.851	+8.7%	0.823	0.913	+10.9%

Boosted EMOS scores 0.9125 and EMOS 1.0097. On EUPPBench NN-Hybrid-Res[†] (0.9134), the DRN (0.9170), SAMOS-MLP (0.9198) and NN-UNK (0.9205) lead; GNN-Geo (0.9468) is the weakest network, ahead of Boosted EMOS (0.9625). The DRN improves on Boosted EMOS by 5.3% (German) and 4.7% (EUPPBench) and GNN-Geo by 3.1% and 1.6%; all four differences are BH-significant within the descriptive family (Table S5), and the DRN is better in each of the seven random folds on both datasets. GNN-Geo is significantly worse than the DRN on both datasets (by 2.3% and 3.2%).

The stage-specified primary tests of NN-Hybrid (10-seed ensembles; Table S2, Section S1) give a mixed picture. On German random folds NN-Hybrid is significantly better than EMOS, Boosted EMOS, NN-UNK, NN-kNN and NN-NoEmb, while its difference from NN-Attr is unresolved by this test; on EUPPBench random folds it is better than both EMOS baselines but significantly worse than NN-UNK (+1.4%) and NN-kNN (+0.7%). Under the claim rule, which also requires the same sign on spatial folds, NN-Hybrid is better than EMOS, Boosted EMOS and NN-kNN on German

Table S5 Diebold–Mariano tests on daily-mean CRPS differences (A – B; negative = A better; 3-seed ensembles), Newey–West lag 5, 95% moving-block bootstrap CI (7-day blocks), Benjamini–Hochberg (BH) at $q = 0.05$. The DRN/GNN-Geo comparisons are *descriptive* (pre-specified pairs, BH within dataset $\times$ fold type, no hypothesis specified); pilot tests use BH over the 6 stage-specified tests per dataset. The tests treat the station set as fixed and ignore seed variance (Section 6.3 (main article)).

Data	Folds	A vs B	Δ	rel. %	DM p	95% CI	BH
German	random	GNN-Geo vs NN-UNK	+0.0134	+1.54	0.0021	[+0.0064, +0.0220]	yes
German	random	GNN-Geo vs NN-kNN	-0.0009	-0.11	0.9	[-0.0158, +0.0149]	no
German	random	GNN-Geo vs Boosted EMOS	-0.0285	-3.13	1.8e-06	[-0.0376, -0.0154]	yes
German	random	GNN-Geo vs DRN	+0.0195	+2.26	4.2e-13	[+0.0150, +0.0253]	yes
German	random	DRN vs NN-NoEmb	+0.0065	+0.75	0.021	[+0.0015, +0.0120]	yes
German	random	DRN vs NN-UNK	-0.0061	-0.70	0.055	[-0.0117, +0.0004]	no
German	random	DRN vs Boosted EMOS	-0.0480	-5.26	$< 10^{-15}$	[-0.0552, -0.0383]	yes
German	random	SAMOS-Attn vs NN-UNK	+0.0032	+0.37	0.24	[-0.0013, +0.0092]	no
German	random	SAMOS-Attn vs NN-kNN	-0.0111	-1.25	0.14	[-0.0261, +0.0042]	no
German	random	SAMOS-Attn vs Boosted EMOS	-0.0387	-4.24	2.9e-12	[-0.0480, -0.0265]	yes
German	spatial	GNN-Geo vs NN-UNK	+0.0125	+1.41	0.013	[+0.0037, +0.0232]	yes
German	spatial	GNN-Geo vs NN-kNN	-0.0023	-0.25	0.8	[-0.0206, +0.0166]	no
German	spatial	GNN-Geo vs Boosted EMOS	-0.0110	-1.20	0.057	[-0.0202, +0.0019]	no
German	spatial	GNN-Geo vs DRN	+0.0253	+2.88	$< 10^{-15}$	[+0.0202, +0.0311]	yes
German	spatial	DRN vs NN-NoEmb	+0.0018	+0.20	0.53	[-0.0033, +0.0074]	no
German	spatial	DRN vs NN-UNK	-0.0128	-1.43	0.0012	[-0.0198, -0.0045]	yes
German	spatial	DRN vs Boosted EMOS	-0.0363	-3.96	1.9e-14	[-0.0441, -0.0262]	yes
German	spatial	SAMOS-Attn vs NN-UNK	-0.0057	-0.64	0.1	[-0.0122, +0.0016]	no
German	spatial	SAMOS-Attn vs NN-kNN	-0.0205	-2.26	0.011	[-0.0372, -0.0042]	yes
German	spatial	SAMOS-Attn vs Boosted EMOS	-0.0293	-3.20	1e-08	[-0.0380, -0.0184]	yes
EUPPBench	random	GNN-Geo vs NN-UNK	+0.0263	+2.85	$< 10^{-15}$	[+0.0229, +0.0299]	yes
EUPPBench	random	GNN-Geo vs NN-kNN	+0.0192	+2.07	6.2e-15	[+0.0143, +0.0241]	yes
EUPPBench	random	GNN-Geo vs Boosted EMOS	-0.0157	-1.63	3.6e-06	[-0.0226, -0.0086]	yes
EUPPBench	random	GNN-Geo vs DRN	+0.0298	+3.25	$< 10^{-15}$	[+0.0269, +0.0330]	yes
EUPPBench	random	DRN vs NN-NoEmb	-0.0257	-2.72	$< 10^{-15}$	[-0.0292, -0.0219]	yes
EUPPBench	random	DRN vs NN-UNK	-0.0035	-0.38	0.022	[-0.0067, -0.0006]	yes
EUPPBench	random	DRN vs Boosted EMOS	-0.0455	-4.73	$< 10^{-15}$	[-0.0534, -0.0379]	yes
EUPPBench	random	SAMOS-Attn vs NN-UNK	+0.0159	+1.72	1.1e-07	[+0.0101, +0.0226]	yes
EUPPBench	random	SAMOS-Attn vs NN-kNN	+0.0088	+0.95	0.0036	[+0.0031, +0.0154]	yes
EUPPBench	random	SAMOS-Attn vs Boosted EMOS	-0.0261	-2.71	1.2e-07	[-0.0359, -0.0158]	yes
EUPPBench	spatial	GNN-Geo vs NN-UNK	+0.0591	+6.02	$< 10^{-15}$	[+0.0519, +0.0674]	yes
EUPPBench	spatial	GNN-Geo vs NN-kNN	+0.0622	+6.35	$< 10^{-15}$	[+0.0550, +0.0706]	yes
EUPPBench	spatial	GNN-Geo vs Boosted EMOS	+0.0653	+6.69	$< 10^{-15}$	[+0.0554, +0.0758]	yes
EUPPBench	spatial	GNN-Geo vs DRN	-0.0364	-3.37	$< 10^{-15}$	[-0.0439, -0.0281]	yes
EUPPBench	spatial	DRN vs NN-NoEmb	+0.0028	+0.26	0.62	[-0.0085, +0.0140]	no
EUPPBench	spatial	DRN vs NN-UNK	+0.0955	+9.71	$< 10^{-15}$	[+0.0874, +0.1039]	yes
EUPPBench	spatial	DRN vs Boosted EMOS	+0.1017	+10.41	$< 10^{-15}$	[+0.0907, +0.1128]	yes
EUPPBench	spatial	SAMOS-Attn vs NN-UNK	+0.0387	+3.93	1.6e-09	[+0.0261, +0.0530]	yes
EUPPBench	spatial	SAMOS-Attn vs NN-kNN	+0.0417	+4.26	8.3e-11	[+0.0292, +0.0560]	yes
EUPPBench	spatial	SAMOS-Attn vs Boosted EMOS	+0.0448	+4.59	1.2e-08	[+0.0296, +0.0621]	yes

data, and better than EMOS and NN-Attr but worse than NN-UNK and NN-kNN on EUPPBench. Because the German result is not independent of the preliminary experiment (Section 4.4 (main article)), the EUPPBench outcome is the relevant replication: the proposal did not replicate as the best new-station method.

Calibration differs between model families (Table S1). Networks with $\sigma = |o_2|$ are under-dispersive at new stations on random folds (80% coverage 0.75–0.79, spread–skill 0.86–0.94), whereas the DRN and GNN-Geo, which use dropout and larger learning rates, are close to nominal (0.81–0.85, 0.98–1.06); Boosted EMOS is slightly over-dispersive on EUPPBench (spread–skill 1.07). These differences are associations with the training set-up; we did not test their cause.

S2.3 Spatially blocked folds

Holding out whole regions affects the two datasets very differently (Fig. S1). On the dense German network the CRPS of every method increases by at most 6.6% relative to random folds, and NN-NoEmb (0.8777 °C), SAMOS-MLP (0.8784) and the DRN (0.8795) have the three lowest pooled three-seed scores, within 0.0018 °C. The DRN remains 4.0% better than Boosted EMOS (0.9158; BH-significant, better in all seven folds and at 86% of the stations), whereas GNN-Geo is 1.2% better ($p = 0.057$). In the most isolated fifth of the German stations (more than about 118km from any training station), the DRN (0.913), NN-NoEmb (0.915) and SAMOS-MLP (0.917) have lower pooled scores than Boosted EMOS (0.931), which has a similar pooled score to NN-kNN (0.932) and NN-UNK (0.936) and a lower score than the hybrid variants (0.963–0.982).

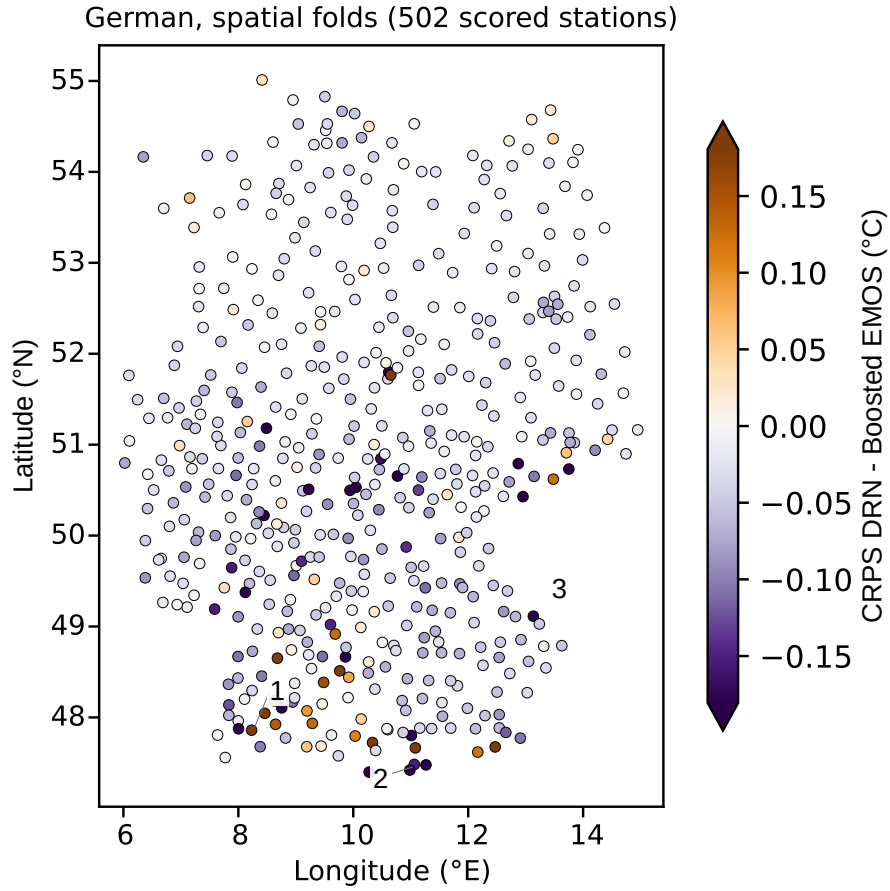
In the original three-seed comparison on EUPPBench, Boosted EMOS has the lowest pooled CRPS (0.9770 °C), followed by NN-kNN (0.9801), NN-Hybrid-Res[†] (0.9803) and NN-UNK (0.9832); these four pooled scores lie within 0.0062 °C. This three-seed ranking is descriptive; seed-nested station inference and ten-seed spatial scores are

reported in Section 5.2 (main article). The remaining methods have higher pooled scores: SAMOS-MLP 1.0023, SAMOS-Attn 1.0218, GNN-Geo 1.0423, NN-NoEmb 1.0759, the DRN 1.0787, NN-Hybrid-kNN[†] 1.1119, NN-Hybrid 1.1131 and NN-Attr 1.1560, worse even than EMOS (1.1395). The DRN is 10.4% worse than Boosted EMOS (difference +0.102 °C, 95% CI [+0.091, +0.113]) and GNN-Geo 6.7% worse; both differences are BH-significant under the daily-mean test. Relative to random folds, the CRPS of Boosted EMOS increases by 1.5%, that of NN-kNN, NN-UNK and NN-Hybrid-Res[†] by 5.7–7.3%, and that of GNN-Geo, NN-NoEmb, the DRN, NN-Hybrid and NN-Attr by 10.1–23.1%.

The pooled DRN deficit is not a uniform degradation. The DRN beats Boosted EMOS in only three of the seven folds, but it has the lower station-mean CRPS at 83 of the 117 stations (71%). Its deficit comes from two folds: fold 2, three French stations (DRN 1.944 vs. 1.243 °C), and fold 5, 17 stations including all four Austrian ones (1.837 vs. 1.169 °C); the fold-wise differences are +0.70 and +0.67 °C there and between −0.08 and +0.03 °C in the other five folds. GNN-Geo, NN-kNN, NN-UNK and NN-Hybrid-Res[†] each beat Boosted EMOS in four of seven folds. With seven folds of very unequal size and three seeds (the DRN’s single-seed CRPS on these folds ranges from 1.002 to 1.179 °C, Table S3), the daily-mean test does not represent uncertainty from the selection of stations, regions or training seeds. The following supplementary analysis addresses these sources and does not establish a robust ranking reversal.

The spatially blocked EUPPBench folds are also where calibration breaks down for the fragile methods: NN-Hybrid covers 73.9% of observations with its central 80% interval (spread–skill 0.79) and the DRN has a mean bias of −0.35 °C, while NN-Hybrid-Res[†] (coverage 0.781, spread–skill 0.95) and Boosted EMOS (0.838, 1.19) retain coverage closer to nominal. In the most isolated fifth of the stations (more than 142 km from any training station), Boosted EMOS (1.114) is best among all methods, followed by NN-kNN (1.203) and NN-Hybrid-Res[†] (1.208); the DRN scores 1.719.

599 Of the five stage-specified EUPPBench replication hypotheses, two were supported:
600 NN-NoEmb beats NN-Attr and NN-Hybrid on spatial folds (H3), and NN-Hybrid
601 beats NN-NoEmb on seen stations (H5). Not supported were H1 (every network beats
602 both EMOS baselines on both fold types), H2 (hybrid, attribute and no-embedding
603 networks within 1% of each other and better than NN-UNK on random folds) and H4
604 (NN-kNN worse than NN-UNK on both fold types; it is worse on random but better
605 on spatial folds).
606
607
608
609



641 **Fig. S2 Exploratory, post hoc.** German DRN minus Boosted EMOS per-station CRPS differences
642 on spatially blocked folds (three-seed ensembles). Orange: first method worse; purple: first method
643 better. The colour scale is clipped at the 95th percentile of $|\Delta|$. Numbered call-outs identify the three
644 largest absolute differences: 1: +0.84; 2: -0.65; 3: -0.48 °C.

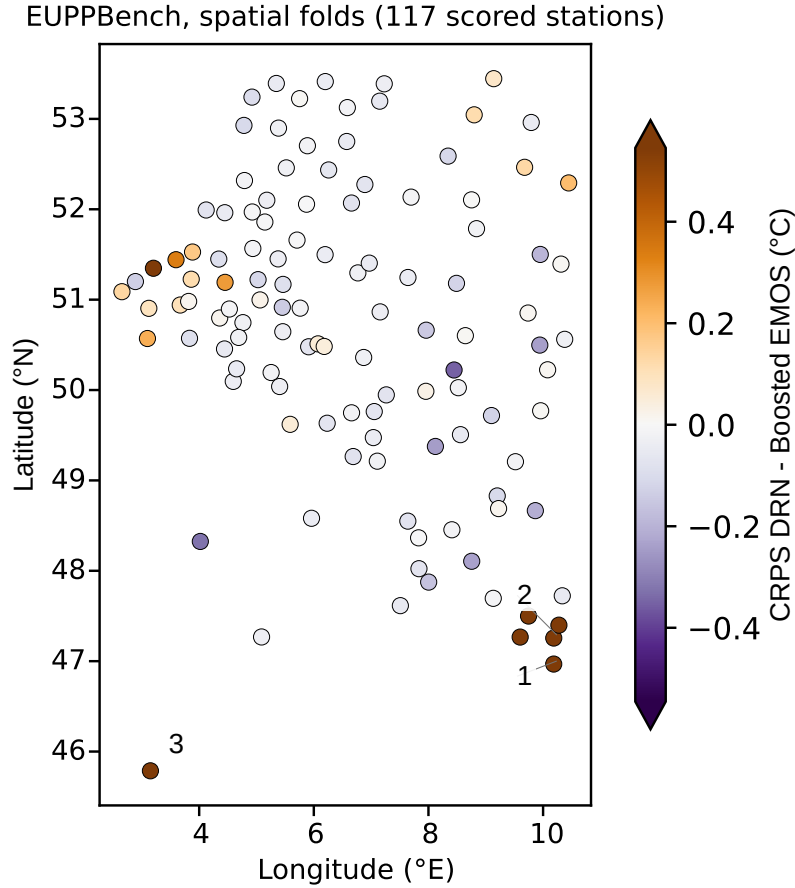


Fig. S3 Exploratory, post hoc. EUPPBench DRN minus Boosted EMOS per-station CRPS differences on spatially blocked folds (three-seed ensembles). Orange: first method worse; purple: first method better. The colour scale is clipped at the 95th percentile of $|\Delta|$. Numbered call-outs identify the three largest absolute differences: 1: Galtür +5.04; 2: Warth +3.75; 3: Clermont-Ferrand +2.47 °C.

S2.4 Stage-specified pilot: a negative result

The pilot combined SAMOS-style standardisation with attention over the eight nearest training stations (Section S5). It failed its go criterion on both datasets (Table S8). On German spatial folds its CRPS was 0.0057 °C below that of the best primary comparator (NN-UNK), but not significantly ($p = 0.10$), and on random folds it was 0.0032 °C worse ($p = 0.24$). On EUPPBench it was significantly worse than the

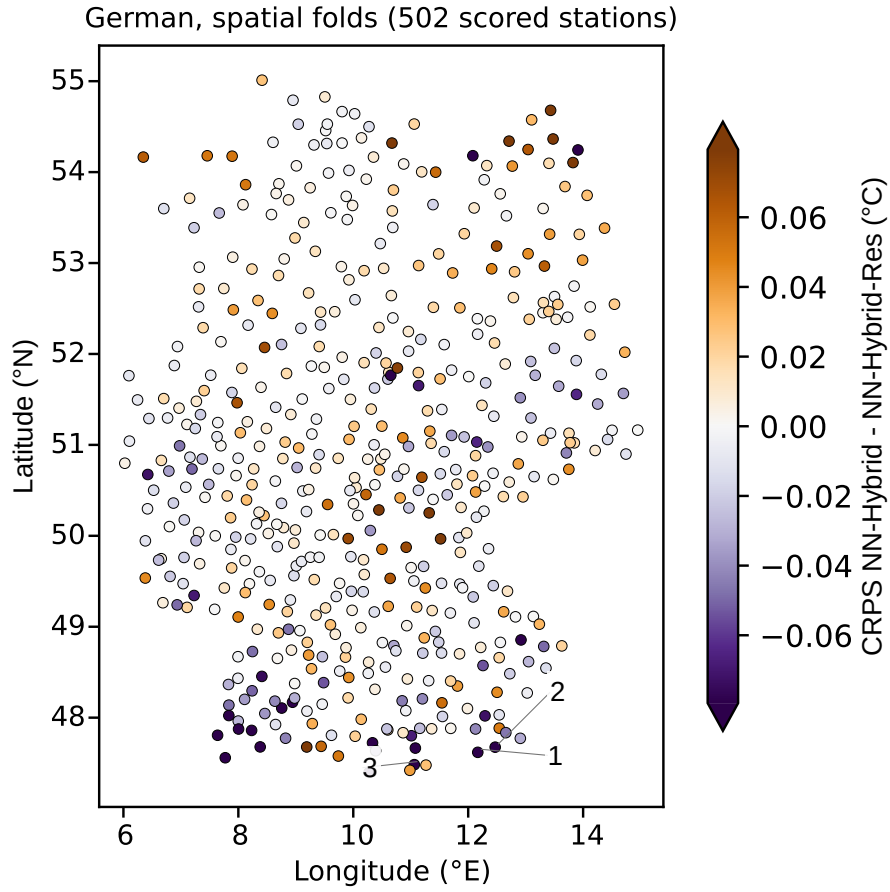


Fig. S4 Exploratory, post hoc. German NN-Hybrid minus NN-Hybrid-Res per-station CRPS differences on spatially blocked folds (three-seed ensembles). Orange: first method worse; purple: first method better. The colour scale is clipped at the 95th percentile of $|\Delta|$. Numbered call-outs identify the three largest absolute differences: 1: -0.37 ; 2: -0.35 ; 3: -0.32 °C.

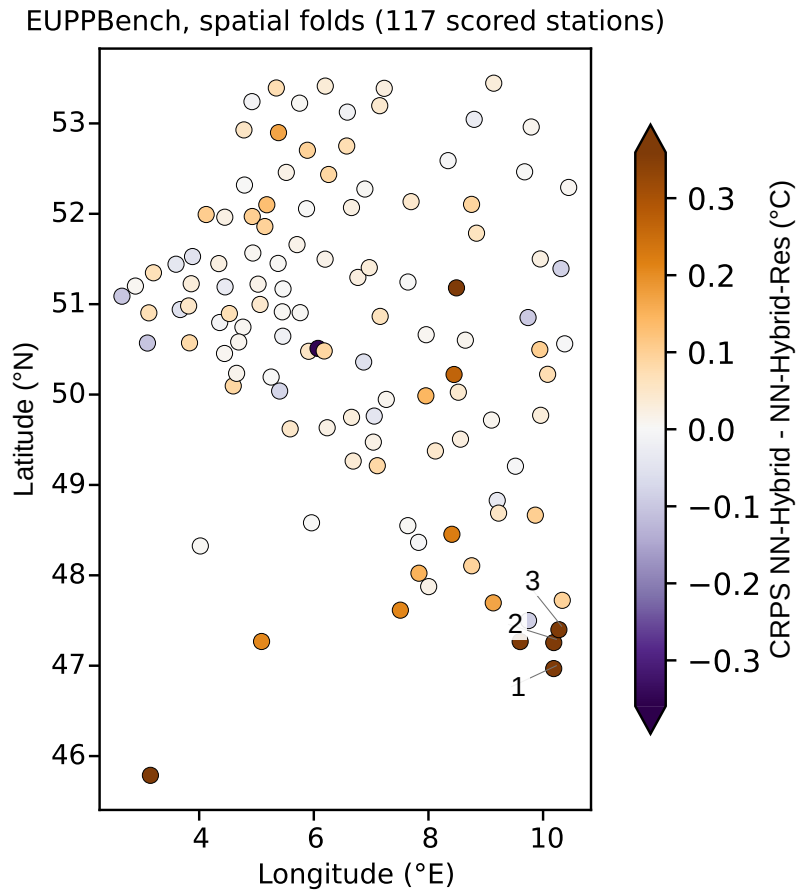


Fig. S5 Exploratory, post hoc. EUPPBench NN-Hybrid minus NN-Hybrid-Res per-station CRPS differences on spatially blocked folds (three-seed ensembles). Orange: first method worse; purple: first method better. The colour scale is clipped at the 95th percentile of $|\Delta|$. Numbered call-outs identify the three largest absolute differences: 1: Galtür +5.46; 2: Warth +2.26; 3: Oberstdorf +1.80 °C.

best comparator on both fold types (+0.0159 and +0.0448 °C). The ablation without neighbours, SAMOS-MLP, had a lower CRPS than SAMOS-Attn in all four dataset and fold combinations.

S3 Exploratory station diagnostics

S3.1 Where do the spatial-fold losses come from? (exploratory, post hoc)

Table S6 Exploratory, post hoc. Spatial folds: CRPS (°C, 3-seed ensembles) of held-out stations whose five attributes (lat, lon, alt, orog, alt−orog) all lie inside the [min, max] range of the fold’s training stations (“in”) or not (“out”). n = number of stations. Bold: lowest in the column; in-range differences between the best methods are small and mostly untested (Section 5.3 (main article)).

Method	German		EUPPBench	
	in ($n=473$)	out ($n=29$)	in ($n=96$)	out ($n=21$)
EMOS	0.996	1.288	1.038	1.599
Boosted EMOS	0.911	0.989	0.954	1.083
NN-UNK	0.888	0.963	0.934	1.207
NN-kNN	0.904	0.953	0.934	1.190
NN-NoEmb	0.867	1.047	0.981	1.505
NN-Attr	0.890	1.284	0.995	1.885
NN-Hybrid	0.883	1.106	0.975	1.740
NN-Hybrid-kNN [†]	0.892	1.143	0.974	1.736
NN-Hybrid-Res [†]	0.882	1.156	0.940	1.164
DRN	0.874	0.966	0.924	1.781
GNN-Geo	0.895	1.069	0.969	1.375
SAMOS-MLP	0.872	0.985	0.930	1.331
SAMOS-Attn	0.876	1.066	0.926	1.458

Status and definition.

This analysis was not part of any stage-specified hypothesis. A held-out station is *out of range* if at least one of five attributes – latitude, longitude, station altitude,

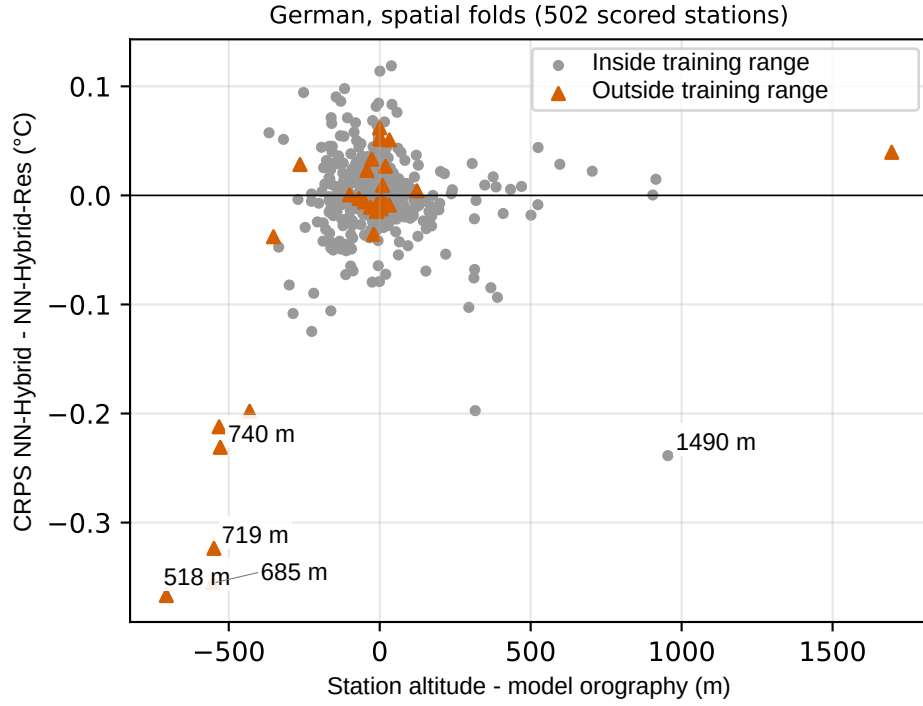


Fig. S6 Exploratory, post hoc. German NN-Hybrid minus NN-Hybrid-Res per-station CRPS difference against station altitude minus model orography on spatially blocked folds (three-seed ensembles). Triangles mark at least one station attribute outside the training range. The five largest $|\Delta|$ are labelled with station names (EUPPBench) or altitude (German). The original plot coordinates and values are preserved.

model orography, or altitude minus orography – lies outside the $[\min, \max]$ interval of that attribute over the training stations of its fold. This axis-aligned box ignores combinations of attributes and the distance outside the range. It was defined and implemented in the diagnosis of the NN-Hybrid results, i.e. with knowledge of which stations had failed for those networks, but before the DRN, GNN-Geo and pilot were specified and trained. The DRN and GNN-Geo rows of Table S6 are therefore out-of-sample with respect to the definition, but no hypothesis about them was specified and the subset tests are uncorrected (Table S7, Section S4). Interpretations in this subsection are exploratory hypotheses. On spatial folds, 21 of the 117 EUPPBench stations (18.1% of the test rows) and 29 of the 502 German stations (5.8%) are out of

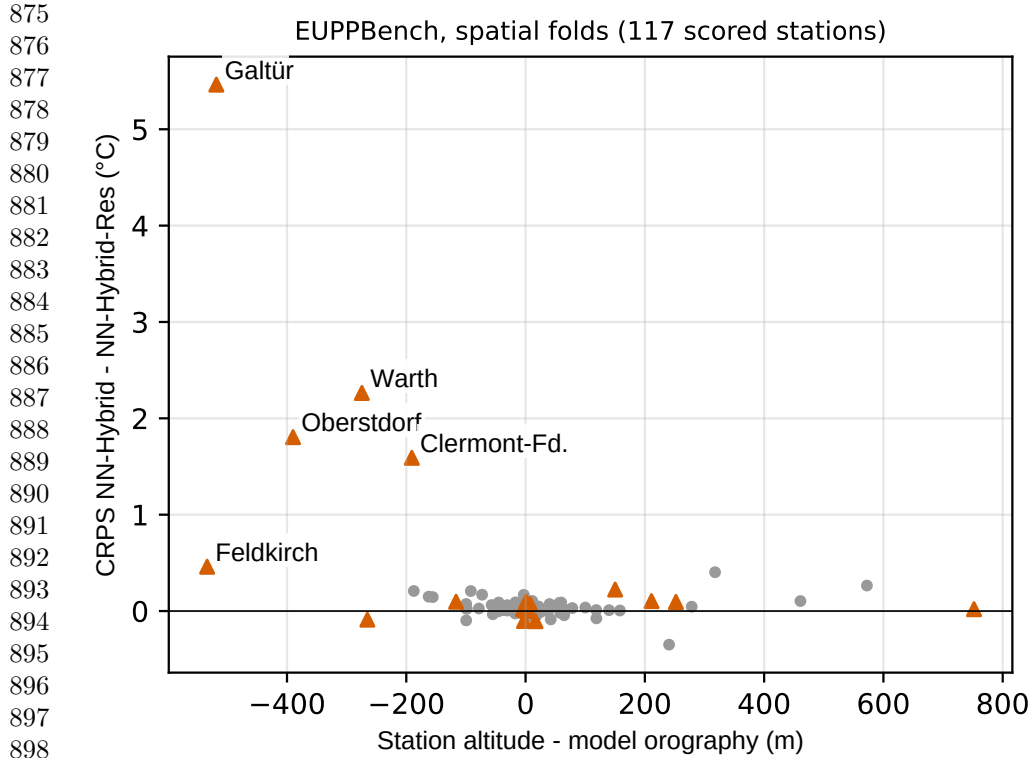


Fig. S7 Exploratory, post hoc. EUPPBench NN-Hybrid minus NN-Hybrid-Res per-station CRPS difference against station altitude minus model orography on spatially blocked folds (three-seed ensembles). Triangles mark at least one station attribute outside the training range. The five largest $|\Delta|$ are labelled with station names (EUPPBench) or altitude (German). The original plot coordinates and values are preserved.

range; on EUPPBench 13 of the 21 are flagged by model orography. On random folds only 9 EUPPBench and 7 German stations are out of range.

In range and out of range.

On the 96 in-range EUPPBench stations the DRN (0.924 °C) and SAMOS-Attn (0.926) have the lowest CRPS, a difference that is not significant (uncorrected $p = 0.48$), followed by SAMOS-MLP (0.930), NN-UNK and NN-kNN (0.934), NN-Hybrid-Res[†] (0.940) and Boosted EMOS (0.954). Within this subset the DRN is 3.15% better than Boosted EMOS (uncorrected $p < 10^{-12}$) and better at 77% of the stations. On the 21 out-of-range stations it scores 1.781 °C against 1.083 for Boosted EMOS (+64%),

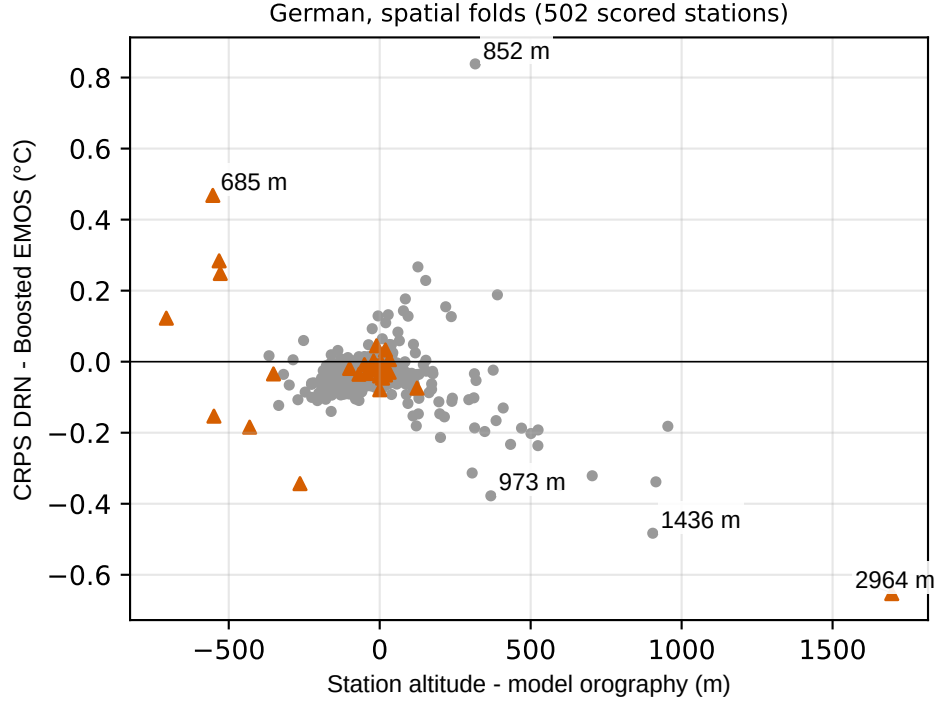


Fig. S8 Exploratory, post hoc. German DRN minus Boosted EMOS per-station CRPS difference against station altitude minus model orography on spatially blocked folds (three-seed ensembles). Triangles mark at least one station attribute outside the training range. The five largest $|\Delta|$ are labelled with station names (EUPPBench) or altitude (German). The original plot coordinates and values are preserved.

yet it is worse there at only 12 of the 21 stations: a few very large errors dominate. The out-of-range stations contribute $+0.126^{\circ}\text{C}$ to the pooled DRN–Boosted-EMOS difference of $+0.102^{\circ}\text{C}$, the in-range stations -0.025°C . GNN-Geo shows the same pattern in weaker form, but is also worse than Boosted EMOS in range ($+0.015$ in, $+0.292^{\circ}\text{C}$ out). On the German spatial folds the DRN is better than Boosted EMOS both in range (-0.037) and out of range (-0.023 , $p = 0.053$).

Which stations fail.

The largest losses of the DRN and of NN-Hybrid occur at the same few stations (Figs. S3 and S5): the Austrian stations Galtür, Warth and Feldkirch, valley stations 518, 274 and 534m below the model orography (the training stations of their fold

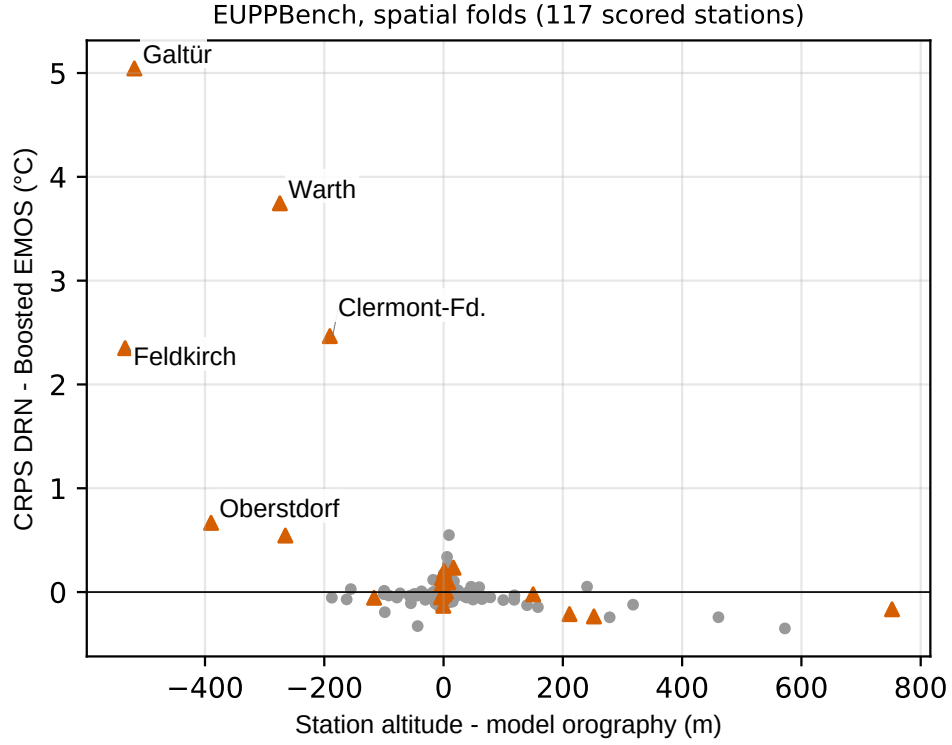


Fig. S9 Exploratory, post hoc. EUPPBench DRN minus Boosted EMOS per-station CRPS difference against station altitude minus model orography on spatially blocked folds (three-seed ensembles). Triangles mark at least one station attribute outside the training range. The five largest $|\Delta|$ are labelled with station names (EUPPBench) or altitude (German). The original plot coordinates and values are preserved.

reach at most 191 m below it), the French station Clermont-Ferrand, 376 km from the nearest training station, and Oberstdorf. At Galtür the raw ensemble is 5.4 °C too cold. Boosted EMOS reduces the station-mean CRPS from 4.77 to 1.65 °C and NN-Hybrid-Res[†] to 1.55 °C (bias +0.05 °C). The networks with an unconstrained mean fail in *both* directions: NN-Hybrid overcorrects to a bias of +8.7 °C (CRPS 7.01), whereas the DRN is even colder than the raw ensemble (bias −8.6 °C, CRPS 6.70). These are the signatures of extrapolation beyond the training attribute range; the stored predictions do not reveal the functional form of the extrapolation.

Residual anchoring.

Replacing only the mean of NN-Hybrid by that of NN-Hybrid-Res[†] reduces the EUPPBench spatial-fold CRPS from 1.1131 to 0.9816 °C, whereas replacing only the standard deviation gives 1.1100; the benefit comes from the anchored mean [Eq. (2) in the main article]. Five stations contribute 75% of the pooled NN-Hybrid minus NN-Hybrid-Res[†] difference (+0.133 °C), and the per-station difference is rank-correlated with attribute novelty ($\rho = 0.26$, $p = 0.004$). The residual head is not uniformly beneficial: on the 29 German out-of-range stations it is *worse* than NN-Hybrid (station-mean difference 0.050 °C) and 16.9% worse than Boosted EMOS, while NN-kNN (0.953), NN-UNK (0.963) and the DRN (0.966) are the best methods there. Further details, including the German Zugspitze counter-example, are given in Section S4.

S4 Exploratory diagnosis: further details

The analyses in this section are exploratory and post hoc.

Table S7 Exploratory, post hoc. Pairwise comparisons within the in-range and out-of-range subsets of Table S6 (spatial folds, 3-seed ensembles). $\Delta = \text{CRPS(A)} - \text{CRPS(B)}$ (negative: A better); DM p on daily means (Newey–West lag 5), *not* corrected for multiplicity and not part of any stage-specified family; “st.” = fraction of stations in the subset where A has the lower station-mean CRPS (descriptive; stations are not independent); “contrib.” = contribution of the subset to the pooled all-station Δ (the two contributions sum to the pooled Δ).

A vs B	pooled Δ	in range				out of range			
		Δ	DM p	st.	contrib.	Δ	DM p	st.	contrib.
<i>German: 473 stations in range, 29 out of range</i>									
DRN vs Boosted EMOS	-0.036	-0.037	7.3e-15	0.87	-0.035	-0.023	0.053	0.72	-0.001
DRN vs SAMOS-Attn	-0.007	-0.001	0.7	0.48	-0.001	-0.099	7.9e-11	0.55	-0.006
DRN vs NN-Hybrid-Res [†]	-0.018	-0.008	0.039	0.44	-0.007	-0.190	< 10 ⁻¹⁵	0.55	-0.011
NN-Hybrid-Res [†] vs Boosted EMOS	-0.018	-0.030	2.1e-07	0.76	-0.028	+0.167	2.8e-10	0.52	+0.010
GNN-Geo vs Boosted EMOS	-0.011	-0.017	0.0035	0.58	-0.016	+0.080	2.8e-05	0.28	+0.005
NN-Hybrid vs NN-Hybrid-Res [†]	-0.002	+0.001	0.34	0.45	+0.001	-0.051	6.3e-06	0.59	-0.003
NN-UNK vs Boosted EMOS	-0.024	-0.023	3.1e-05	0.73	-0.022	-0.026	0.0065	0.69	-0.002
<i>EUPPBench: 96 stations in range, 21 out of range</i>									
DRN vs Boosted EMOS	+0.102	-0.030	8e-13	0.77	-0.025	+0.698	< 10 ⁻¹⁵	0.43	+0.126
DRN vs SAMOS-Attn	+0.057	-0.002	0.48	0.46	-0.002	+0.323	< 10 ⁻¹⁵	0.43	+0.058
DRN vs NN-Hybrid-Res [†]	+0.098	-0.016	1.6e-08	0.50	-0.013	+0.617	< 10 ⁻¹⁵	0.29	+0.112
NN-Hybrid-Res [†] vs Boosted EMOS	+0.003	-0.014	0.003	0.61	-0.011	+0.081	< 10 ⁻¹⁵	0.62	+0.015
GNN-Geo vs Boosted EMOS	+0.065	+0.015	0.00015	0.69	+0.012	+0.292	< 10 ⁻¹⁵	0.29	+0.053
NN-Hybrid vs NN-Hybrid-Res [†]	+0.133	+0.035	< 10 ⁻¹⁵	0.21	+0.029	+0.576	< 10 ⁻¹⁵	0.14	+0.104
NN-UNK vs Boosted EMOS	+0.006	-0.020	1.6e-05	0.71	-0.016	+0.124	< 10 ⁻¹⁵	0.43	+0.022

1059 *Pairwise comparisons within subsets.*

1060
1061 Table S7 gives uncorrected DM tests within the in-range and out-of-range subsets, the
1062 fraction of stations where A is better, and the contribution of each subset to the pooled
1063 difference. On EUPPBench, NN-Hybrid-Res[†] is better than NN-Hybrid also in range
1064 difference. On EUPPBench, NN-Hybrid-Res[†] is better than NN-Hybrid also in range
1065 difference. On EUPPBench, NN-Hybrid-Res[†] is better than NN-Hybrid also in range
1066 difference. On EUPPBench, NN-Hybrid-Res[†] is better than NN-Hybrid also in range
1067 difference. On EUPPBench, NN-Hybrid-Res[†] is better than NN-Hybrid also in range
1068 difference. On EUPPBench, NN-Hybrid-Res[†] is better than NN-Hybrid also in range
1069 difference. On EUPPBench, NN-Hybrid-Res[†] is better than NN-Hybrid also in range
1070 difference. On EUPPBench, NN-Hybrid-Res[†] is better than NN-Hybrid also in range
1071 Distance to the nearest training station correlates only weakly with the per-station
1072 gain of the residual head (EUPPBench $\rho = 0.20$, $p = 0.03$; German $\rho = -0.13$).

1074

1075 *Station details.*

1076
1077 At the three Austrian valley stations the raw ensemble is 5.4 (Galtür), 4.6 (Warth) and
1078 5.6 °C (Feldkirch) too cold. At Galtür the station-mean CRPS is 3.41 for NN-UNK,
1079 4.15 for NN-NoEmb, 7.01 for NN-Hybrid and 8.30 °C for NN-Attr, with biases between
1080 +4.0 and +9.9 °C; the DRN has a bias of −8.6 °C, and GNN-Geo, which aggregates
1081 the forecasts of nearby stations, −2.8 °C (CRPS 2.35). At Warth and Feldkirch the
1082 DRN bias is −7.2 and −4.9 °C. Per station, the DRN minus Boosted EMOS difference
1083 is +5.04 °C at Galtür and NN-Hybrid minus NN-Hybrid-Res[†] +5.46 °C.
1084
1085
1086
1087
1088
1089

1090 *A counter-example.*

1091
1092 Extrapolation is not harmful per se. The most extreme German station, the Zugspitze
1093 summit (2964 m, 1696 m above the model orography), lies far outside the altitude
1094 range of its training fold, yet NN-Hybrid and NN-Hybrid-Res[†] are similarly good there
1095 (1.19 and 1.15 °C) and the DRN is 0.65 °C better than Boosted EMOS. German train-
1096 ing folds contain stations up to 954 m above the model orography, so the correction
1097 is extrapolated in a direction that the training data span.
1098
1099
1100
1101
1102
1103
1104

S5 Stage-specified pilot: SAMOS standardisation	1105
with neighbour attention	1106
	1107
	1108
	1109
Motivation and planning	1110
	1111
The exploratory diagnosis of Section 5.3 (main article) suggested that new-station	1112
errors arise from unconstrained extrapolation in station-attribute space. Before any	1113
pilot run we specified a method that (i) models climatologically standardised anoma-	1114
lies, so that the network does not have to learn station climatology, and (ii) borrows	1115
information from nearby <i>training</i> stations instead of extrapolating from attributes,	1116
together with a go/no-go rule. A two-epoch smoke test on tiny subsamples printed test-	1117
period CRPS values of untrained models (German 1.04/1.06, EUPPBench 1.82/1.85	1118
for SAMOS-Attn/SAMOS-MLP; recorded in the pilot protocol notes, raw output not	1119
saved); no design choice was changed after this output.	1120
	1121
	1122
	1123
	1124
	1125
	1126
	1127
Method	1128
	1129
1. <i>SAMOS-style standardisation</i> (Dabernig et al. 2017; Stauffer et al. 2017), fitted on	1130
training stations and the fitting period only: ridge regressions give climatological	1131
means and log-variances of the observation and of the ensemble-mean t2m, with	1132
regressors $[1, \text{lat}, \text{lon}, \text{alt}, \text{alt} - \text{orog}] \times [1, \text{first and second annual harmonics}]$ (plus	1133
lead one-hot $\times [1, \sin, \cos]$ on EUPPBench). This yields m_y, s_y, m_f, s_f for every	1134
station and time, including held-out stations.	1135
	1136
2. <i>Output in standardised space</i> , anchored at the forecast anomaly $z_f = (m - m_f)/s_f$:	1137
$\mu = m_y + s_y(z_f + g)$ and $\sigma = s_y \sqrt{\text{softplus}(c)^2 + \text{softplus}(d)^2 z_s^2}$, where $z_s = s/s_f$ is	1138
the ensemble spread scaled by the forecast climatological standard deviation and	1139
g, c, d are network outputs.	1140
	1141
3. <i>Attention over the $K = 8$ nearest available training stations</i> (target excluded)	1142
plus a learned null slot. Each neighbour token holds the neighbour’s standardised	1143
	1144
	1145
	1146
	1147
	1148
	1149
	1150

1151 forecast and spread, its anomaly difference to the target, the distance, the altitude
 1152 and alt – orog differences, and a learned 4-dimensional embedding of that training
 1153 station, the only channel through which observation-derived station information
 1154 enters. The target’s own embedding is never used.
 1155
 1156
 1157 4. *Random station masking* during training to create pseudo-new stations: per batch
 1158 a fraction $U(0, 0.5)$ of training stations is removed from the neighbour pool; per
 1159 sample, with probability 0.5, all training stations within $r \sim U(0, 300)$ km of the
 1160 target are removed.
 1161
 1162
 1163 5. Hidden size 128, attention dimension 32, Adam (learning rate 10^{-3}), batch 4096,
 1164 at most 30 epochs, patience 3; nothing was tuned.
 1165
 1166
 1167 The ablation SAMOS-MLP uses the same standardisation and output without
 1168 neighbours, attention or masking.
 1169
 1170
 1171
 1172 **Decision rule**
 1173
 1174 Primary comparators: NN-UNK, NN-kNN and Boosted EMOS (3-seed ensembles).
 1175
 1176 GO required, on German data, (1) a BH-significant improvement over the best primary
 1177 comparator on spatial folds with a bootstrap CI excluding zero and (2) a lower CRPS
 1178 than the best primary comparator on random folds. BH correction is over the six
 1179 primary tests per dataset. Phase 2 (EUPPBench) was to run only if the German
 1180 spatial point estimate was better than the best comparator, and was then judged by
 1181 the same rule as a replication check. The automatic Phase-1 check crashed because
 1182 of an evaluator bug, and the launcher consequently skipped Phase 2; after the bug
 1183 was fixed the evaluation was rerun on the unchanged Phase-1 predictions and Phase
 1184
 1185
 1186
 1187
 1188
 1189
 1190 2 was started as specified.
 1191
 1192
 1193
 1194
 1195
 1196

Table S8 Stage-specified pilot (SAMOS-Attn) against the best of the three primary comparators on each fold type (3-seed ensembles). $\Delta = \text{CRPS}(\text{SAMOS-Attn}) - \text{CRPS}(\text{best})$; negative favours the pilot. GO required a BH-significant spatial improvement with bootstrap CI excluding 0 and a same-sign random-fold improvement. Last column: single-seed CRPS of SAMOS-Attn for seeds 0/1/2.

Data	Folds	best comparator	CRPS best	CRPS pilot	Δ	DM p	BH	seeds 0/1/2
German	random	NN-UNK	0.8706	0.8738	+0.0032	0.24	no	0.896/0.899/0.891
German	spatial	NN-UNK	0.8922	0.8865	-0.0057	0.1	no	0.910/0.903/0.905
Decision: NO-GO								
EUPPBench	random	NN-UNK	0.9205	0.9364	+0.0159	1.1e-07	yes	0.945/0.944/0.979
EUPPBench	spatial	Boosted EMOS	0.9770	1.0218	+0.0448	1.2e-08	yes	0.984/1.177/0.991
Decision: NO-GO								

Results

- German: the spatial point estimate favoured the pilot (-0.0057 , about 0.6%) but was not significant ($p = 0.10$, CI $[-0.0122, +0.0016]$), and the random-fold difference had the wrong sign ($+0.0032$, $p = 0.24$): **NO-GO**. Because the spatial point estimate was below the best comparator, Phase 2 ran as specified.
- EUPPBench: significantly worse than the best primary comparator on both fold types: **NO-GO**.
- Seeds: the single-seed CRPS of SAMOS-Attn varies strongly on EUPPBench spatial folds (0.984, 1.177 and 0.991 °C for seeds 0–2; Table S8), so the 3-seed ensemble result there is uncertain; the decision on German data, where seeds agree closely, does not depend on this.
- Ablation (descriptive): SAMOS-MLP beats SAMOS-Attn everywhere (German random 0.8544 vs. 0.8738, spatial 0.8784 vs. 0.8865; EUPPBench random 0.9198 vs. 0.9364, spatial 1.0023 vs. 1.0218). The neighbour attention therefore did not help; plain standardisation was competitive on German data but lost to Boosted EMOS, NN-UNK, NN-kNN and NN-Hybrid-Res[†] on EUPPBench spatial folds.
- Following the specified rule, the method is not promoted and no further tuning was done.

1243 **S6 Supplementary robustness analyses**

1244

1245 This section reports the completed supplementary stage under the internal working
1246 plan; Section [S6.5](#) adds post hoc intervals computed from the stored predictions. The
1247 earlier three-seed tables and exploratory diagnostics are retained; the tables below
1248 label ensemble sizes explicitly. Stage 5 was specified after the original results were
1249 known and is a robustness analysis of the same data, not an independent replication.
1250
1251

1252

1253 **S6.1 Supplementary robustness: relative deterioration without** 1254 **an established ranking reversal**

1255

1256 With ten-seed spatial ensembles, EUPPBench DiD is $+0.151^{\circ}\text{C}$ for the DRN
1257 (seed-nested station 95% interval $[+0.041, +0.317]$) and $+0.102^{\circ}\text{C}$ for GNN-Geo
1258 ($[+0.037, +0.183]$; Table 3 (main article)). Both point estimates are positive in all six
1259 additional spatial partitions with matched three-seed ensembles (Table [S9](#)). The pre-
1260 stated C1 condition is therefore met: in this EUPPBench comparison, random station
1261 holdouts yield a more favourable relative comparison of these flexible networks with
1262 Boosted EMOS than spatially blocked holdouts. This is an overstatement of relative
1263 cold-start performance only when the spatial-holdout target is taken as the reference.
1264 German DiD intervals include zero for both methods. This meets the historical descrip-
1265 tive C3 condition. With matched three-seed ensembles, however, both German DiD
1266 intervals exclude zero (Section [S6.5](#)), so meeting C3 depends on the unequal primary
1267 ensemble sizes. A difference in statistical significance is not evidence of a significant
1268 difference between datasets. No direct cross-dataset contrast interval was computed.
1269 Network density is confounded with geographic domain, forecast systems, periods,
1270 lead times and predictors, so these data cannot establish a density effect.
1271

1272 Neither method meets C2’s stronger ranking-reversal condition. The DRN’s EUPP-
1273 Bench spatial difference from Boosted EMOS is $+0.105^{\circ}\text{C}$ with station interval
1274 $[-0.006, +0.275]$, so its random-fold advantage is no longer established on spatial folds.
1275
1276
1277
1278

GNN-Geo has a spatial deficit of $+0.087^{\circ}\text{C}$ (station interval $[+0.018, +0.186]$ and BH rejection), but its random-fold advantage has interval $[-0.033, +0.021]$ and is itself uncertain. Thus GNN-Geo is worse on average over EUPPBench spatial stations under C4, while the DRN comparison is unresolved; absence of a detected difference is not evidence of equivalence.

The within-spatial-design fold-block intervals for model differences include zero for both DRN ($[-0.029, +0.393]^{\circ}\text{C}$) and GNN-Geo ($[-0.039, +0.327]^{\circ}\text{C}$). Large deficits are regionally concentrated: the DRN scores 1.885 versus 1.243°C for Boosted EMOS in fold 2 (three French stations) and 1.884 versus 1.169 in fold 5 (17 stations including Austria). GNN-Geo’s largest deficit is also in fold 5 (1.826 versus 1.169). The pooled, record-weighted GNN deficit therefore does not establish a consistent disadvantage across regions. These intervals concern GNN or DRN versus Boosted EMOS within spatial folds, not the DiD between spatial and random designs; paired regional block intervals for the DiD are given in Section S6.5. On German spatial folds, both DRN (0.8759°C) and GNN-Geo (0.8934) improve on Boosted EMOS (0.9158) under C4, with fold-block intervals also below zero.

The ten-seed EUPPBench spatial scores of NN-Hybrid-Res[†] (0.9744°C), NN-UNK (0.9760), Boosted EMOS (0.9770) and NN-kNN (0.9804) are close (Table 2 (main article)). In particular, neither NN-Hybrid-Res nor NN-UNK has an established ordering against Boosted EMOS under C4–C5; the three-seed statement that Boosted EMOS has the lowest pooled score does not carry over numerically. The added EMOS-RegIDW and SAMOS-Lin score 1.0178 and 1.0165°C on EUPPBench spatial folds (0.9646 and 0.9579 on German folds). Both have higher CRPS than Boosted EMOS under C4 on both datasets, and neither meets C6’s condition of outperforming every EUPPBench network. Sections S6.2–S6.5 report the additional partitions, stopping sensitivity, forecast-only checks, and matched-seed and regional-block DiD intervals; the full spatial table is Table 2 (main article).

S6.2 Ten-seed spatial comparisons and additional partitions

Table 2 (main article) gives final primary spatial scores for every method. Table S9 gives the DRN and GNN DiD under six extra EUPPBench partitions. All twelve point estimates and their recorded station-and-seed intervals are positive. The historical C1 condition only required positive point estimates in at least four partitions in addition to the primary-partition station intervals; meeting it does not establish independent or region-level replication. These repeated partitions are correlated because they reuse observations. Leave-one-country-out folds contain 4, 27, 49, 7 and 30 stations (Austria, Belgium, Germany, France and the Netherlands); k-means folds also differ substantially in size.

Table S9 DiD relative to Boosted EMOS under the six additional EUPPBench spatial partitions. All network ensembles in these comparisons use 3 seeds. Units: °C; intervals are seed-nested station-bootstrap 95% intervals. These partitions share observations and are not independent replications.

Partition	DRN	95% CI	GNN-Geo	95% CI
$K = 7$, seed 2027	+0.206	[+0.038, +0.647]	+0.095	[+0.051, +0.159]
$K = 7$, seed 2028	+0.112	[+0.033, +0.232]	+0.055	[+0.017, +0.166]
$K = 7$, seed 2029	+0.140	[+0.035, +0.315]	+0.052	[+0.017, +0.101]
$K = 5$, seed 2026	+0.141	[+0.042, +0.303]	+0.099	[+0.044, +0.172]
$K = 10$, seed 2026	+0.094	[+0.029, +0.182]	+0.104	[+0.035, +0.277]
Leave one country out	+0.101	[+0.064, +0.151]	+0.131	[+0.062, +0.234]

S6.3 Boosted-EMOS stopping sensitivity

Table S10 reports the specified stopping sensitivity. More iterations slightly improve random-fold CRPS, but spatial CRPS and dispersion deteriorate beyond 500–1000 iterations. AIC-type selection reaches iteration 4000 in 66 of 70 random/spatial fold/lead fits, and 3978–3994 in the remaining four; validation-selected stopping gives spatial CRPS 1.0015 °C and spread–skill 3.747, compared with 0.9770 and 1.187 for the primary 1000-iteration setting. Validation uses seen training stations, so its preference need not improve transfer. The 500-iteration spatial test score is 0.9750 °C, but

selecting it from test performance would violate the protocol; the primary setting is unchanged.

Table S10 EUPPBench Boosted-EMOS stopping sensitivity. CRPS is in $^{\circ}\text{C}$; the spatial spread-skill ratio (SSR) is also shown. The specified 1000-iteration setting remains primary; test scores are not used to choose the stopping point. AIC and validation select separately within each fold and lead.

Stopping	Random CRPS	Spatial CRPS	Spatial SSR
100	0.99831	1.01516	1.096
250	0.96848	0.97930	1.043
500	0.96414	0.97500	1.083
1000	0.96250	0.97700	1.187
2000	0.96153	0.98542	1.839
4000	0.96109	1.00207	3.744
AIC	0.96108	1.00207	3.744
Validation	0.96110	1.00148	3.747

S6.4 Forecast-only neighbour availability

The original GNN and attention models used neighbours only when both a forecast and an observation were available. The sensitivity restores forecast features for included EUPPBench stations with missing observations (86 368 fit/validation rows and 3960 test rows); observation values remain unavailable and these rows do not enter the loss or the verification sample. All test forecasts on the included station/initialisation/lead grid are available. Neighbour sets change in 10 910 of 423 090 scored GNN cases (2.58%) on both random and spatial folds, and in 27 370 spatial SAMOS-Attn cases (6.47%). The GNN fraction is saved in the original run logs. The SAMOS-Attn fraction was reconstructed from the test availability metadata and the eight-nearest-training- station rule; the same reconstruction reproduces the GNN log counts exactly.

All comparisons in Table S11 use three-seed ensembles. Spatial changes are -0.0140°C for GNN-Geo and $+0.0138^{\circ}\text{C}$ for SAMOS-Attn, exceeding the specified

absolute 0.005 °C reporting threshold. Both are shown alongside the historical imple-
mentations. The station-and-seed intervals for their score changes include zero, which
does not establish equivalence or remove the information-condition limitation. The
forecast-only GNN’s three-seed DiD is +0.0705 °C, with station-and-seed interval
[+0.0247, +0.1747] °C (Section S6.5). The random GNN change is −0.0035 °C. The
original predictions remain primary. These reruns change neighbour availability dur-
ing fitting as well as prediction, so score changes cannot be attributed solely to the
percentage of changed test neighbours. In the supplied German files, forecasts are
stored only where observations exist, preventing an analogous recovery of forecast-only
neighbours.

Table S11 EUPPBench sensitivity to defining neighbour availability using forecasts only. Three-seed ensembles are compared with their original three-seed versions. Δ is corrected minus original CRPS (°C); intervals include station and seed resampling. Historical results retain their reported status; the information-condition correction is shown separately.

Folds	Method	Original	Corrected	Δ	95% CI
Spatial	GNN-Geo	1.04234	1.02829	-0.0140	[-0.083, +0.062]
Spatial	SAMOS-Attn	1.02185	1.03568	+0.0138	[-0.180, +0.232]
Random	GNN-Geo	0.94678	0.94328	-0.0035	[-0.027, +0.015]

S6.5 Matched-seed, forecast-only and regional-block DiD intervals

These post hoc analyses use only preserved per-seed predictions; no model was
retrained. They reuse the primary seed-nested station bootstrap (2000 replicates) with
the same station weights and seed-resampling streams; the primary DiD intervals of
Table 3 (main article) are reproduced exactly. Matched contrasts restrict the spatial
ensembles to seeds 0–2, so both designs use three-seed ensembles (Table S12).

On EUPPBench the matched DRN DiD is $+0.1472^{\circ}\text{C}$ $[+0.0433, +0.3383]^{\circ}\text{C}$ and the matched GNN-Geo DiD is $+0.0811^{\circ}\text{C}$ $[+0.0333, +0.1714]^{\circ}\text{C}$. The forecast-only GNN-Geo DiD is $+0.0705^{\circ}\text{C}$ $[+0.0247, +0.1747]^{\circ}\text{C}$. All three station-and-seed intervals exclude zero, so in these stored predictions the direction of relative deterioration under spatial blocking does not depend on the ensemble-size mismatch or on observation-based neighbour selection. The intervals overlap substantially with the primary ones and do not separate the magnitudes of the configurations.

On the German data the matched DiDs are DRN $+0.0117^{\circ}\text{C}$ $[+0.0040, +0.0194]^{\circ}\text{C}$ (primary $+0.0081^{\circ}\text{C}$ $[-0.0010, +0.0121]^{\circ}\text{C}$) and GNN-Geo $+0.0175^{\circ}\text{C}$ $[+0.0033, +0.0307]^{\circ}\text{C}$ (primary $+0.0062^{\circ}\text{C}$ $[-0.0089, +0.0145]^{\circ}\text{C}$). Both matched German intervals exclude zero, whereas both primary ones include it. The German finding that the DiD intervals include zero therefore depends on comparing ten spatial with three random seeds; with matched three-seed ensembles a small positive German DiD is also detected. This further weakens any contrast between the two datasets that rests on one significant and one nonsignificant interval.

Tables S13 and S14 replace station resampling with a paired regional block bootstrap: whole clusters of stations are resampled and the same cluster weights are applied to both designs, nested with the seed resampling. Clusters are the seven primary spatial folds, the five countries (EUPPBench) and k -means regions on station coordinates. For EUPPBench, block intervals are 1.2–3.1 times as wide as the corresponding station intervals; 23 of 25 exclude zero, the exceptions being GNN-Geo (primary), primary folds: $[-0.0031, +0.3117]^{\circ}\text{C}$; GNN-Geo (primary), k -means, $G=10$: $[-0.0053, +0.2628]^{\circ}\text{C}$. Several lower bounds lie close to zero, and with five to 30 unequal clusters these percentile intervals are approximate. The regional analysis therefore supports the sign of the EUPPBench DiD less firmly than station resampling and leaves its magnitude poorly determined. German block intervals are 1.0–1.6 times as wide as the station intervals; 8 of 8 primary German block intervals include

zero, and 7 of 8 matched German block intervals exclude zero, the exception being GNN-Geo (matched), primary folds: $[-0.0011, +0.0386]$ °C.

Table S12 Post hoc DiD intervals relative to Boosted EMOS (°C). Intervals use the primary seed-nested station bootstrap (2000 replicates; the same shared station weights and seed-resampling streams as the primary analysis). Seeds 10/3 is the primary configuration; 3/3 rows restrict spatial ensembles to seeds 0–2 (matched). Primary rows reproduce Table 3 (main article). FO: forecast-only neighbour availability. P is the bootstrap fraction of replicates with DiD > 0.

Data	Method	Seeds sp/rd	DiD	95% CI	P
EUPPBench	DRN	10/3	+0.1506	[+0.0411, +0.3174]	1.000
EUPPBench	GNN-Geo	10/3	+0.1022	[+0.0372, +0.1827]	0.999
EUPPBench	DRN	3/3	+0.1472	[+0.0433, +0.3383]	1.000
EUPPBench	GNN-Geo	3/3	+0.0811	[+0.0333, +0.1714]	0.999
EUPPBench	GNN-Geo FO	3/3	+0.0705	[+0.0247, +0.1747]	0.999
German	DRN	10/3	+0.0081	[−0.0010, +0.0121]	0.961
German	GNN-Geo	10/3	+0.0062	[−0.0089, +0.0145]	0.788
German	DRN	3/3	+0.0117	[+0.0040, +0.0194]	0.998
German	GNN-Geo	3/3	+0.0175	[+0.0033, +0.0307]	0.989

Table S13 Paired regional block bootstrap of the EUPPBench DiD relative to Boosted EMOS (°C). In each of 2000 replicates whole clusters of stations are resampled with replacement and the same cluster weights are applied to the spatial and random designs, nested with the corresponding seed resampling. G is the number of clusters; k -means regions use the repository's station-coordinate k -means (seed 2026). FO: forecast-only neighbour availability. P is the bootstrap fraction of replicates with DiD > 0. Bootstrap intervals with few, unequal clusters are approximate.

Contrast	Resampling unit	G	95% CI	P
DRN, seeds 10/3 (primary); DiD +0.1506				
	Stations (primary)	117	[+0.0411, +0.3174]	1.000
	Primary spatial folds	7	[+0.0183, +0.4376]	1.000
	Countries	5	[+0.0237, +0.7931]	1.000
	k -means regions	10	[+0.0156, +0.4089]	0.999
	k -means regions	20	[+0.0122, +0.4455]	1.000
	k -means regions	30	[+0.0131, +0.4670]	0.999
GNN-Geo, seeds 10/3 (primary); DiD +0.1022				
	Stations (primary)	117	[+0.0372, +0.1827]	0.999
	Primary spatial folds	7	[−0.0031, +0.3117]	0.961
	Countries	5	[+0.0203, +0.4697]	0.999
	k -means regions	10	[−0.0053, +0.2628]	0.958
	k -means regions	20	[+0.0077, +0.2577]	0.992
	k -means regions	30	[+0.0067, +0.2553]	0.989
DRN, seeds 3/3 (matched); DiD +0.1472				
	Stations (primary)	117	[+0.0433, +0.3383]	1.000
	Primary spatial folds	7	[+0.0214, +0.4434]	1.000
	Countries	5	[+0.0245, +0.8195]	1.000
	k -means regions	10	[+0.0168, +0.4033]	0.999
	k -means regions	20	[+0.0135, +0.4705]	0.999
	k -means regions	30	[+0.0156, +0.4340]	1.000
GNN-Geo, seeds 3/3 (matched); DiD +0.0811				
	Stations (primary)	117	[+0.0333, +0.1714]	0.999
	Primary spatial folds	7	[+0.0039, +0.2366]	0.988
	Countries	5	[+0.0211, +0.3319]	0.999
	k -means regions	10	[+0.0006, +0.2144]	0.979
	k -means regions	20	[+0.0106, +0.2192]	0.996
	k -means regions	30	[+0.0112, +0.2107]	0.994
GNN-Geo FO, seeds 3/3 (matched); DiD +0.0705				
	Stations (primary)	117	[+0.0247, +0.1747]	0.999
	Primary spatial folds	7	[+0.0001, +0.2233]	0.976
	Countries	5	[+0.0208, +0.2818]	0.999
	k -means regions	10	[+0.0002, +0.2081]	0.975
	k -means regions	20	[+0.0135, +0.1998]	0.999
	k -means regions	30	[+0.0061, +0.1919]	0.990

Table S14 Paired regional block bootstrap of the German DiD relative to Boosted EMOS ($^{\circ}\text{C}$). In each of 2000 replicates whole clusters of stations are resampled with replacement and the same cluster weights are applied to the spatial and random designs, nested with the corresponding seed resampling. G is the number of clusters; k -means regions use the repository's station-coordinate k -means (seed 2026). P is the bootstrap fraction of replicates with $\text{DiD} > 0$. Bootstrap intervals with few, unequal clusters are approximate.

Contrast	Resampling unit	G	95% CI	P
DRN, seeds 10/3 (primary); DiD +0.0081				
	Stations (primary)	502	$[-0.0010, +0.0121]$	0.961
	Primary spatial folds	7	$[-0.0016, +0.0150]$	0.940
	k -means regions	20	$[-0.0016, +0.0152]$	0.949
	k -means regions	50	$[-0.0018, +0.0146]$	0.946
	k -means regions	100	$[-0.0011, +0.0134]$	0.961
GNN-Geo, seeds 10/3 (primary); DiD +0.0062				
	Stations (primary)	502	$[-0.0089, +0.0145]$	0.788
	Primary spatial folds	7	$[-0.0099, +0.0174]$	0.755
	k -means regions	20	$[-0.0100, +0.0148]$	0.771
	k -means regions	50	$[-0.0085, +0.0141]$	0.793
	k -means regions	100	$[-0.0092, +0.0146]$	0.774
DRN, seeds 3/3 (matched); DiD +0.0117				
	Stations (primary)	502	$[+0.0040, +0.0194]$	0.998
	Primary spatial folds	7	$[+0.0010, +0.0242]$	0.987
	k -means regions	20	$[+0.0011, +0.0251]$	0.988
	k -means regions	50	$[+0.0016, +0.0241]$	0.986
	k -means regions	100	$[+0.0030, +0.0217]$	0.996
GNN-Geo, seeds 3/3 (matched); DiD +0.0175				
	Stations (primary)	502	$[+0.0033, +0.0307]$	0.989
	Primary spatial folds	7	$[-0.0011, +0.0386]$	0.968
	k -means regions	20	$[+0.0024, +0.0330]$	0.987
	k -means regions	50	$[+0.0041, +0.0326]$	0.990
	k -means regions	100	$[+0.0034, +0.0311]$	0.991

S6.6 Reproducibility and remaining sensitivities

For the ten-seed EUPPBench spatial DRN and GNN comparisons against Boosted EMOS, automatic Newey–West bandwidths are 17 and 19 days (Andrews: 29 and 30), compared with the original fixed lag of 5. Both daily-series tests reject under every reported bandwidth, the primary BH correction and the fixed- b sensitivity. The 30-day moving-block intervals are $[0.092, 0.122]^{\circ}\text{C}$ (DRN) and $[0.072, 0.106]^{\circ}\text{C}$ (GNN), still entirely positive. Thus lengthening the temporal dependence correction does not account for the loss of significance of the DRN comparison under station-and-seed

resampling; these procedures quantify different sources of uncertainty. On German spatial folds, both networks' differences remain negative with all three daily block lengths and all reported DM bandwidths. These sensitivities concern fixed fitted ensembles on the observed station network and do not replace the station or region intervals.

The final evaluation preserves seed-nested and fixed-ensemble station intervals, seed-nested fold intervals, daily-series DM tests and bandwidth/block-length sensitivities, per-seed scores, and bootstrap ranking fractions. The claim rules combine the specified station interval and multiplicity-adjusted DM test; favourable individual sensitivities are not substituted for either requirement. Within-design region intervals retain a separate interpretive role, as described in Section 5.2 (main article). They cannot substitute for the paired regional DiD intervals of Section S6.5. The main article's comparator-stopping point estimates remain arithmetic sensitivities of stored summaries without intervals; the matched-ensemble and forecast-only contrasts have their own intervals in Section S6.5, and the primary intervals must not be transferred to them.

Nested validation at held-out stations, a paired same-station seen/new comparison, matched predictor sets, buffered folds, a variogram analysis and alternative attribute-novelty definitions were outside this stage. The new partition and seed analyses do not resolve those limitations. The original exploratory out-of-range diagnostics are not recomputed with ten seeds and must not be interpreted as ten-seed estimates.

1703 References

- 1704
1705 Andrews DWK (1991) Heteroskedasticity and autocorrelation consistent covariance
1706 matrix estimation. *Econometrica* 59(3):817–858. <https://doi.org/10.2307/2938229>
1707
1708
1709 Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practi-
1710 cal and powerful approach to multiple testing. *Journal of the Royal Statistical*
1711 *Society Series B: Statistical Methodology* 57(1):289–300. [https://doi.org/10.1111/j.](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x)
1712 [2517-6161.1995.tb02031.x](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x)
1713
1714
1715
1716
1717 Dabernig M, Mayr GJ, Messner JW, et al (2017) Spatial ensemble post-processing
1718 with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*
1719 143(703):909–916. <https://doi.org/10.1002/qj.2975>
1720
1721
1722
1723 Diebold FX, Mariano RS (1995) Comparing predictive accuracy. *Journal of Busi-*
1724 *ness & Economic Statistics* 13(3):253–263. [https://doi.org/10.1080/07350015.1995.](https://doi.org/10.1080/07350015.1995.10524599)
1725 [10524599](https://doi.org/10.1080/07350015.1995.10524599)
1726
1727
1728
1729 Harvey D, Leybourne S, Newbold P (1997) Testing the equality of prediction mean
1730 squared errors. *International Journal of Forecasting* 13(2):281–291. [https://doi.org/](https://doi.org/10.1016/S0169-2070(96)00719-4)
1731 [10.1016/S0169-2070\(96\)00719-4](https://doi.org/10.1016/S0169-2070(96)00719-4)
1732
1733
1734 Kiefer NM, Vogelsang TJ (2005) A new asymptotic theory for heteroskedasticity-
1735 autocorrelation robust tests. *Econometric Theory* 21(6):1130–1164. [https://doi.org/](https://doi.org/10.1017/S0266466605050565)
1736 [10.1017/S0266466605050565](https://doi.org/10.1017/S0266466605050565)
1737
1738
1739
1740 Künsch HR (1989) The jackknife and the bootstrap for general stationary observations.
1741 *The Annals of Statistics* 17(3):1217–1241. <https://doi.org/10.1214/aos/1176347265>
1742
1743
1744
1745
1746
1747
1748

Lakatos M (2026) Statistical versus machine learning-based spatial interpolation of	1749
post-processed ensemble weather forecasts. arXiv preprint arXiv:2609.07512, https://doi.org/10.48550/arXiv.2609.07512	1750 1751 1752 1753 1754
Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and	1755
autocorrelation consistent covariance matrix. <i>Econometrica</i> 55(3):703–708. https://doi.org/10.2307/1913610	1756 1757 1758 1759 1760
Newey WK, West KD (1994) Automatic lag selection in covariance matrix estimation.	1761
The Review of Economic Studies 61(4):631–653. https://doi.org/10.2307/2297912	1762 1763 1764
Rasp S, Lerch S (2018) Neural networks for postprocessing ensemble weather	1765
forecasts. <i>Monthly Weather Review</i> 146(11):3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1	1766 1767 1768 1769 1770
Stauffer R, Umlauf N, Messner JW, et al (2017) Ensemble postprocessing of daily	1771
precipitation sums over complex terrain using censored high-resolution standard-	1772
ized anomalies. <i>Monthly Weather Review</i> 145(3):955–969. https://doi.org/10.1175/MWR-D-16-0260.1	1773 1774 1775 1776 1777 1778 1779 1780 1781 1782 1783 1784 1785 1786 1787 1788 1789 1790 1791 1792 1793 1794