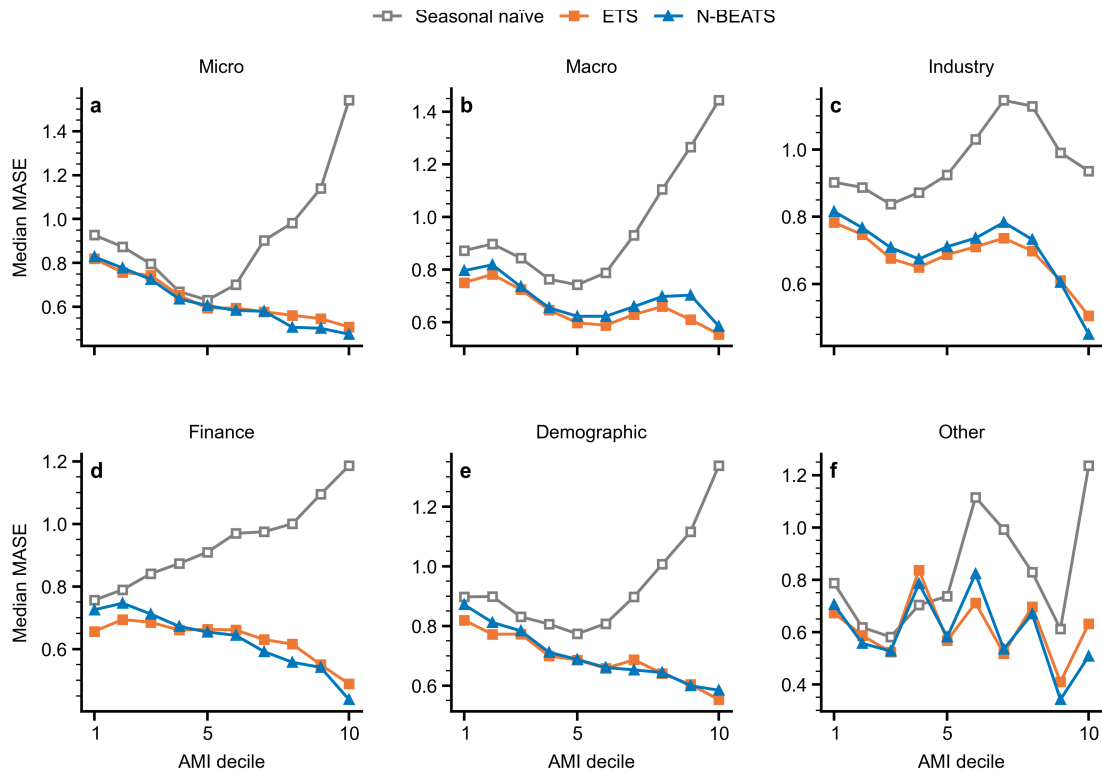


Supplementary Information for: Horizon-resolved Forecastability of Time Series via Auto Mutual Information

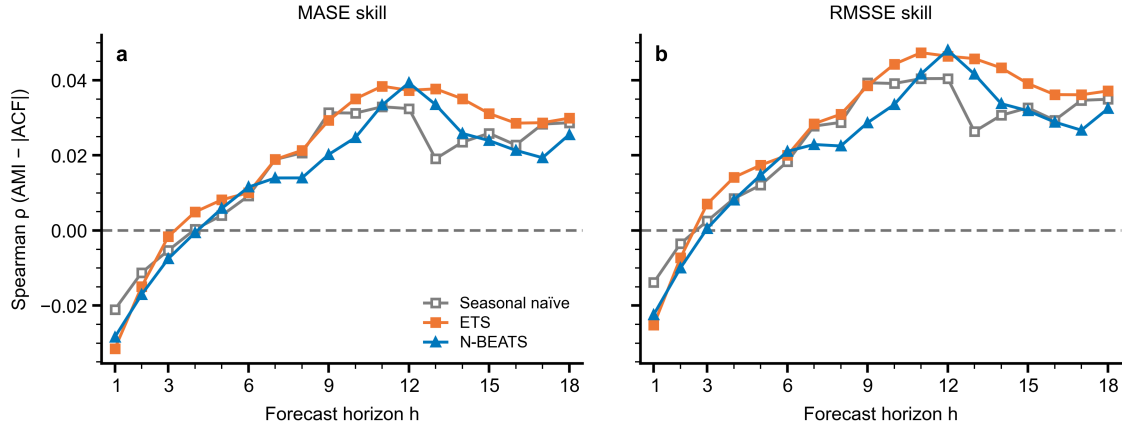
Peter M. Catt
Virtual Blue Ltd, Auckland, New Zealand
peter.catt@virtualblue.co.nz

Supplementary Figures

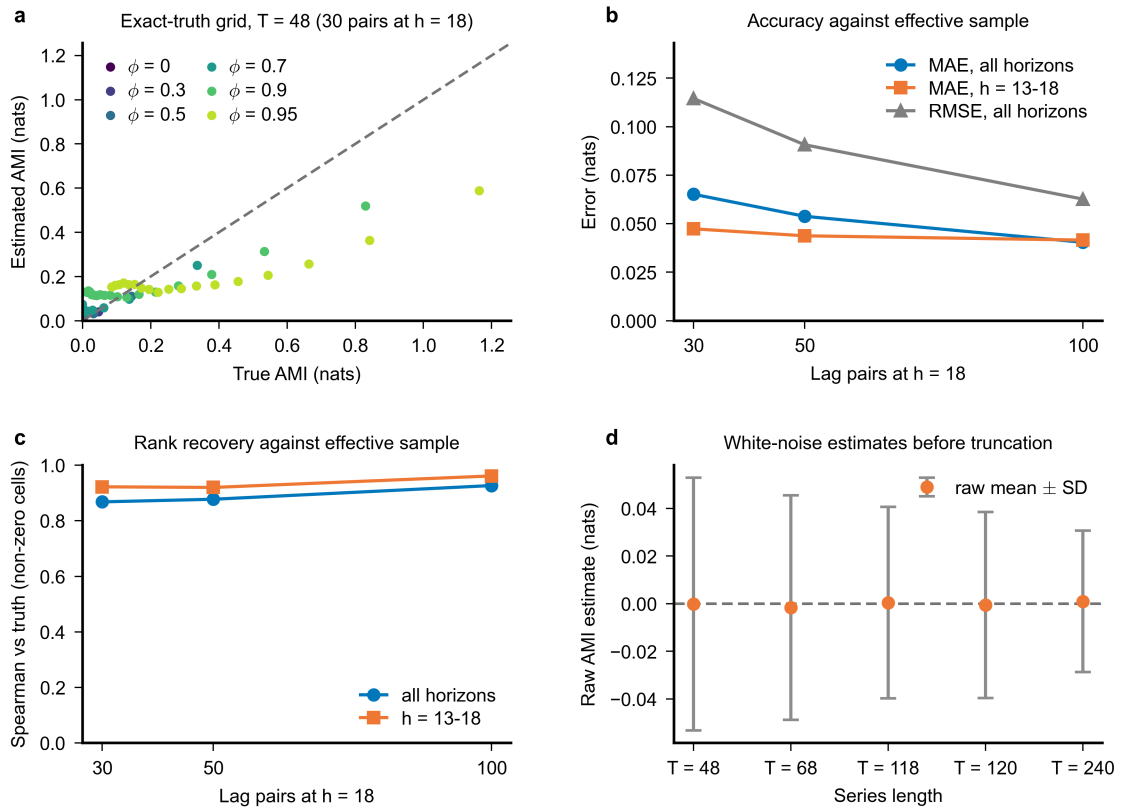
Supplementary Fig. S1. Decile gradient by M4 Monthly category. Median MASE by within-horizon AMI decile (1 = lowest AMI, 10 = highest) for seasonal naïve, ETS, and N-BEATS, shown separately for each of the six M4 Monthly categories (mean of per-horizon decile medians within category). The within-category pattern mirrors the pooled decile gradient reported in the main text. The Other panel rests on 269 series (266 for N-BEATS) and is correspondingly noisy.



Supplementary Fig. S2. AMI-over- $|\text{ACF}|$ rank-association advantage by horizon under both error metrics. Spearman correlation difference (AMI minus $|\text{ACF}|$, against realised skill) by forecast horizon for seasonal naïve, ETS, and N-BEATS, under (a) MASE and (b) RMSSE skill. The advantage concentrates at longer horizons under both metrics, so the horizon pattern reported in the main text is not metric-specific.



Supplementary Fig. S3. Small-sample estimator validation at 30 to 100 lag pairs. Exact-truth AR(1) grid rerun at series lengths $T = 48, 68$, and 118 (30, 50, and 100 pairs at $h = 18$; master seed 20260720). (a) Estimated against true AMI at $T = 48$. (b) Mean absolute error and RMSE against the pair count at $h = 18$, overall and for $h = 13-18$. (c) Spearman correlation with truth (non-zero cells) against the pair count. (d) White-noise estimates before truncation (mean $\pm$ SD of the raw estimates) across series lengths, including the main study's $T = 120$ and $T = 240$.



Supplementary Tables

Supplementary Table S1. AMI estimator sensitivity: horizon-averaged Spearman correlation with MASE skill by neighbour-count specification.

Specification	Seasonal naïve	ETS	N-BEATS
median $\{k = 3, k = 5, k = 10\}$ (main)	−0.1038	0.1121	0.1259
$k = 3$	−0.0976	0.1179	0.1304
$k = 5$	−0.1039	0.1120	0.1259
$k = 10$	−0.1127	0.0995	0.1152
Raw (un-truncated) AMI, median k	−0.1038	0.1120	0.1259

Supplementary Table S2. Length-adjusted diagnostic-skill correlations: mean Spearman ρ across horizons $h = 1$ –18 between each diagnostic and realised skill (−MASE), unadjusted and as partial correlations controlling the rank of training length. 95% confidence intervals from 1,000 bootstrap resamples of series. Differences are computed before rounding.

Model	Unadjusted AMI	Adjusted AMI [95% CI]	Adjusted ACF [95% CI]	Adjusted advantage [95% CI]
ETS	0.1121	0.1300 [0.1246, 0.1361]	0.1055 [0.0999, 0.1115]	+0.0245 [0.0218, 0.0270]
N-BEATS	0.1259	0.1417 [0.1356, 0.1475]	0.1231 [0.1175, 0.1291]	+0.0186 [0.0160, 0.0210]
Seasonal naïve	−0.1038	−0.0625 [−0.0683, −0.0562]	−0.0816 [−0.0871, −0.0755]	+0.0191 [0.0167, 0.0215]

Supplementary Table S3. Diagnostic-skill correlations within training-length quartiles: mean Spearman ρ across horizons between each diagnostic and realised skill, computed within series-level quartiles of training length (quartile boundaries 82, 202, and 306 observations; 12,514, 11,497, 15,970, and 8,011 series respectively). The boundaries fall on heavily populated length values, so the quartile counts are unequal.

Model	Diagnostic	Q1	Q2	Q3	Q4
ETS	AMI	0.1319	0.1384	0.1370	0.0830
ETS	ACF	0.1068	0.1155	0.1080	0.0553
N-BEATS	AMI	0.1732	0.1132	0.1562	0.0884
N-BEATS	ACF	0.1516	0.0998	0.1296	0.0591
Seasonal naïve	AMI	0.0649	−0.0759	−0.0973	−0.1502
Seasonal naïve	ACF	0.0508	−0.0915	−0.1280	−0.1713

Supplementary Table S4. Full N-BEATS training configuration.

Setting	Value
Architecture	Generic (identity basis, no basis expansion)
Stacks / blocks per stack / layers per block	3 / 3 / 4
Hidden units per fully connected layer	512
Input window (lookback)	48 observations (four seasonal cycles)
Forecast horizon	18 (direct multi-step)
Loss	MAE
Optimiser / learning rate	Adam / 10^{-3}
Training steps / batch size	5,000 / 1,024
Per-series normalisation	robust scaling
Random seed	42
Implementation	neuralforecast 3.1.2 (PyTorch 2.9.1), Apple MPS accelerator

Supplementary Table S5. Synthetic estimator-validation designs: the main study and the small-sample extension. Both use the KSG-I estimator with the median over $k \in \{3, 5, 10\}$ truncated at zero, true AMI $-\frac{1}{2} \ln(1 - \phi^{2h})$ for Gaussian AR(1), and per-stream seeding `default_rng((seed, stream, indices))`.

Design element	Main study	Small-sample extension
Master seed	20260714	20260720
AR(1) coefficients ϕ	0, 0.3, 0.5, 0.7, 0.9, 0.95	0, 0.3, 0.5, 0.7, 0.9, 0.95
Horizons	1–18	1–18
Grid series lengths T	240, 120	48, 68, 118
Lag pairs at $h = 18$	222, 102	30, 50, 100
Series per grid cell	20	100
White-noise series lengths	240, 120	48, 68, 118, 120, 240
White-noise series per length	100	100
Seasonal AR(12) check	$\phi = 0.8$, $T = 240$, 40 series	not rerun
Monotone-transform check	$\exp(x)$ and x^3 at $T = 240$	not rerun

Supplementary Table S6. Zero-truncation frequency of the raw AMI estimate across the M4 Monthly panel: share of finite raw estimates truncated at zero (per cent), by training-length band and horizon group.

Training-length band	h 1–6	h 7–12	h 13–18	All horizons
$T < 120$	5.99	8.41	10.19	8.20
$120 \leq T < 240$	1.61	2.54	3.53	2.56
$240 \leq T < 480$	0.12	0.19	0.36	0.22
$T \geq 480$	0.10	0.12	0.36	0.19
All series	2.41	3.47	4.38	3.42

Supplementary Table S7. Robustness slices of the headline diagnostic-skill associations: horizon-averaged Spearman ρ with realised skill ($-MASE$) for AMI and $|ACF|$, and their difference, for the full sample (reference; reproduces Tables 1 and 2 of the main text), for horizons 1–12 only (dropping $h \geq 13$, where zero-truncation concentrates), and excluding the shortest quarter of series (training length ≤ 82 ; 35,478 series retained).

Sample	Model	AMI	$ ACF $	Advantage
Full sample	Seasonal naïve	−0.1038	−0.1200	+0.0162
Full sample	ETS	0.1121	0.0928	+0.0193
Full sample	N-BEATS	0.1259	0.1115	+0.0144
Horizons 1–12 only	Seasonal naïve	−0.0586	−0.0706	+0.0120
Horizons 1–12 only	ETS	0.1625	0.1496	+0.0130
Horizons 1–12 only	N-BEATS	0.1695	0.1604	+0.0091
Excl. shortest quartile	Seasonal naïve	−0.1060	−0.1266	+0.0206
Excl. shortest quartile	ETS	0.1198	0.0961	+0.0237
Excl. shortest quartile	N-BEATS	0.1314	0.1134	+0.0180

Supplementary Table S8. Small-sample estimator accuracy on the exact-truth AR(1) grid. Mean absolute error, bias, and RMSE of the truncated headline estimate (nats) and Spearman correlation with truth over non-zero cells, by series length. The $T = 120$ and $T = 240$ rows restate the main study (20 series per cell, master seed 20260714); the extension rows use 100 series per cell (master seed 20260720). The truncated share is the series-grain share of finite raw estimates truncated at zero on the grid.

T	Pairs at $h = 18$	MAE	Bias	RMSE	ρ vs truth	ρ vs truth, h 13–18	Truncated share
48	30	0.065	−0.010	0.115	0.868	0.922	0.36
68	50	0.054	−0.003	0.091	0.877	0.920	0.34
118	100	0.040	+0.007	0.063	0.927	0.961	0.30
120	102	0.041	+0.006	0.065	0.903	—	—
240	222	0.027	+0.012	0.040	0.928	—	—

Supplementary Note 1: Estimator sensitivity

Results with each individual neighbour count ($k = 3$, $k = 5$, and $k = 10$) are compared with the main specification, the median across the three counts. If the AMI-skill relationship is robust, horizon-averaged Spearman correlations should change only modestly across specifications, and the relative ordering of models should be preserved. Sensitivity to the neighbour count is reported in Supplementary Table S1. Across the $k = 3$, $k = 5$, and $k = 10$ specifications, the horizon-averaged correlations change only modestly: the largest deviation from the main value is about 0.013, and the sign and the model ordering are preserved in every specification, indicating that the AMI-skill association is not an artefact of the neighbour-count choice. The final row of Supplementary Table S1 repeats the analysis without the zero-truncation of negative estimates. Truncation at zero is a monotone non-decreasing map, so a rank correlation can change only through the ties it creates among the truncated cells, which are the lowest-AMI 3.4 per cent of the panel and carry almost no rank information either way; the correlations accordingly agree to within 0.0002 for every probe, confirming that the truncation rule has no material effect.

Supplementary Note 2: Training-length control

Training length T varies from 48 to 2,794 observations across the panel and is a candidate common cause of the diagnostic-skill association: the number of lag pairs available to the KSG estimator is $T - h$, so the AMI estimate depends on T even at fixed population dependence, and longer series could in principle also be forecast more accurately. Three analyses address this. First, the marginal associations with T : training length correlates with both diagnostics (mean Spearman across horizons 0.4574 with AMI and 0.4705 with $|ACF|$) but is essentially uncorrelated with realised skill for ETS (-0.0088) and N-BEATS (-0.0014); the seasonal naïve association (-0.1068) is consistent with the length dependence of the MASE denominator discussed in the main text. Second, partial Spearman correlations controlling the rank of T (invariant to any monotone transform of T , including $\log T$), computed per horizon and averaged, with 95% confidence intervals from 1,000 bootstrap resamples of series (Supplementary Table S2): adjustment does not attenuate the diagnostic-skill association for either forecasting model (the modest increase in the partials is largely the mechanical variance deflation that follows from partialling out a variable correlated with the diagnostic but not the outcome), and the AMI advantage over $|ACF|$ persists, with confidence intervals excluding zero for both models. At long horizons ($h = 13$ – 18), where zero-truncation is most frequent, the adjusted AMI correlation remains positive for both models (ETS 0.0220; N-BEATS 0.0507) whilst the adjusted $|ACF|$ correlation for ETS is negative (-0.0131). Third, an assumption-free cross-check within series-level quartiles of T (boundaries 82, 202, and 306 observations; Supplementary Table S3): the AMI-skill correlation is positive in all four quartiles for both forecasting models, and AMI matches or exceeds $|ACF|$ in all eight quartile-model cells. The seasonal naïve correlation changes sign in the shortest quartile, further evidence that its association with the diagnostics is an artefact of the training-period denominator rather than a diagnostic property. Together these results indicate that training length acts as noise on the diagnostic side rather than as a confounder: controlling its influence leaves the association intact.

Supplementary Note 3: Small-sample estimator validation

The main synthetic study validates the estimator at series lengths $T = 240$ and $T = 120$, which supply at least 102 lag pairs at every horizon. The M4 Monthly panel, however, admits series down to $T = 48$, so estimates at long horizons operate on as few as 30 pairs, and zero-truncation concentrates in exactly those cells (Supplementary Table S6). The small-sample extension (Supplementary Table S8 and Supplementary Fig. S3) closes that gap: the exact-truth AR(1) grid is rerun at $T = 48, 68$, and 118 (30, 50, and 100 pairs at $h = 18$) with 100 series per cell under master seed 20260720. Point accuracy degrades gracefully as the pair count falls (mean absolute error 0.065 nats at 30 pairs against 0.027 at $T = 240$), whilst rank recovery, the property the analysis relies on, remains high: the Spearman correlation with truth over non-zero cells is 0.87 to 0.93 overall and 0.92 to 0.96 when restricted to $h = 13$ – 18 . The white-noise estimates are also reported before truncation: the raw estimates are centred

on zero at every length (mean -0.0002 nats at $T = 48$ and $+0.0010$ at $T = 240$, with 50 to 57 per cent of raw estimates negative), so the near-zero truncated floor reported in the main text reflects an unbiased estimator rather than a property manufactured by the truncation rule. Combined with the robustness slices of Supplementary Table S7, which show the headline associations and the AMI advantage persisting in sign when $h \geq 13$ is excluded and strengthening when the shortest quarter of series is dropped, this settles the reliability question at the panel's smallest effective samples.