Comparator choice in the evaluation of density neural MIDAS nowcasts

Hyun Hak Kim*

August 2026

Abstract

Density neural MIDAS models report sizeable gains for nowcasting US GDP growth, measured against members of their own model family. We ask what happens when the comparison set is not chosen that way. We implement a published QRNN-MIDAS model to its authors' specification, place it in a three-tier comparison set anchored by a quarterly-updating autoregression, and score every arm separately at each monthly origin inside the target quarter, under a pre-specified protocol. The published model performs worse than the autoregression in all eighteen pre-specified headline cells, and so does the structured neural estimator this paper set out to evaluate. Post-hoc analyses then measure what the sample can support. A model confidence set retains the historical mean and the random walk in every cell; at the headline origins the design resolves CRPS differences of roughly twenty per cent of the benchmark's loss, while the best point estimate there is 15.9 per cent; three of 64 target quarters carry seventy to eighty per cent of the squared loss difference behind each competitive comparison; and a stability-targeted fluctuation test rejects nowhere. Defensible design choices — the composition and size of the comparison set, the resolution of the horizon axis, and the loss — each move what the comparison reports. On samples of this size the ranking of density mixed-frequency nowcasts is a joint property of the models and the comparison design.

Keywords: nowcasting; mixed-frequency data; MIDAS; density forecasting; neural networks; forecast evaluation

JEL classification: C45; C52; C53; E37

Running title: Comparator choice in density nowcasts

*Professor, Department of Economics, Kookmin University, Seoul, Korea. hyunhak.kim@kookmin.ac.kr

1 Introduction

Quarterly GDP is published with a lag, and a large body of work asks how much of that lag can be closed with monthly indicators. Giannone et al. (2008) formalised nowcasting as the reading of a real-time data flow, and two families of tools dominate the practice: dynamic factor and bridge models (Bánbura et al., 2013), and mixed-data sampling regressions (Ghysels et al., 2007; Andreou et al., 2010; Foroni et al., 2015), applied to US output growth by Clements and Galvão (2008) and set against mixed-frequency VARs by Kuzin et al. (2011), who find the two approaches complementary across horizons. The monthly panel used here is by now conventional (McCracken and Ng, 2016).

Interest has meanwhile moved from point forecasts toward whole predictive distributions, scored by proper rules (Gneiting and Raftery, 2007) and compared by tests built for densities (Amisano and Giacomini, 2007). Density nowcasts of GDP have been produced by combination across models (Aastveit et al., 2014), and the growth-at-risk agenda of Adrian et al. (2019) supplied the policy audience: when the object of interest is the lower tail of output growth, a density rather than a point is what a nowcast has to deliver.

The natural next step has been to make the mapping from monthly data to that distribution flexible, and the machine-learning evidence on whether that helps is mixed in an instructive way. Richardson et al. (2021) find machine-learning algorithms beating both an autoregression and a factor model for New Zealand GDP in a fully real-time design; Medeiros et al. (2021) report the same for US inflation with random forests; Goulet Coulombe et al. (2022) conclude across a large design space that the gains, where they exist, come from nonlinearity and are concentrated in particular periods — a finding our Section 5 echoes from the opposite direction. Babii et al. (2022) bring formal guarantees to machine-learning MIDAS regressions, while Németh (2026), comparing five neural architectures for US GDP nowcasting on the same monthly database used here, finds the simpler architectures no worse than the ones built for long memory. Within this literature, Xu et al. (2021) combine a quantile regression neural network with a MIDAS lag polynomial and report gains for US GDP growth of roughly 15 to 47 per cent over QR-MIDAS and 8 to 38 per cent over a quantile regression neural network, across five quantile levels.

One feature of the comparison in Xu et al. (2021) motivates this paper. Its columns are QR-MIDAS, QRNN and QRNN-MIDAS: every reported gain is a gain over another member of the neural and MIDAS families that the proposed model combines. There is no autoregression, no random walk, no factor model and no unconditional mean. This is a property of the comparison set, not a defect of the arithmetic, and it raises a question that the published numbers cannot answer on their own — what happens when a benchmark from outside those families is added.

We answer it directly. We implement QRNN-MIDAS to the authors' specification, place it in a comparison set organised by the question each member answers rather than by family resemblance, and score every arm at each monthly origin inside the target quarter, on a release-aware pseudo-real-time panel, under a protocol frozen before the scoring. The horizon axis follows Kim and Swanson (2018); inference is by a block bootstrap over target quarters rather than by the asymptotic test of

Diebold and Mariano (1995), for reasons the data make plain below.

The published model performs worse than a quarterly-updating autoregression in every one of the eighteen pre-specified headline cells — two regimes, three within-quarter origins, three losses — by between 3.7 and 43.5 per cent of the benchmark's loss. So does the structured neural estimator this paper set out to evaluate, and so do two further neural arms. We state this about our own estimator first because it is the fact that makes the rest of the paper readable as evidence rather than as advocacy.

The negative result does not license its converse. A model confidence set (Hansen et al., 2011) retains the historical mean and the random walk in every one of the cells we examine, so no arm here can be said to beat the benchmark either — including the Student-t factor model and the quantile gradient booster whose point estimates are best. The reason is measurable: at 64 target quarters the design resolves differences of roughly twenty per cent of the benchmark's loss at the headline origin, where the largest point estimate is 15.9 per cent; across all eighteen headline cells the factor model's point estimate reaches 28.4 per cent, and no interval clears zero anywhere. Three of those 64 quarters supply seventy to eighty per cent of the squared loss difference behind each competitive comparison, and they are mostly the financial crisis. Removing the crisis quarters changes signs in seven of sixty model-cells, in both directions; the fluctuation test of Giacomini and Rossi (2010) nonetheless detects no instability, because the changes are smaller than the intervals containing them. This sample determines neither the ranking among the retained arms nor the stability of that ranking, and we report that as a measurement of the design's resolution rather than as an apology for it.

Between those two findings sits the argument of Section 4. Four choices made before any model is fitted each move what the comparison reports: which models enter the comparison set, how finely the horizon axis is resolved — pooling the origins reverses the ranking of two arms that origin-resolved scoring separates — how many models are compared at once, since the model confidence set takes a maximum over them, and which loss does the scoring, since the exclusions here occur under CRPS and not under squared error. None of the four is a mistake, and we have made each of them ourselves in one direction or another.

Horizon-resolved evaluation of mixed-frequency forecasts is not new; the axis we use is that of Kim and Swanson (2018). Statistics based on forecast revisions are not new either; they are the multi-horizon restrictions of Patton and Timmermann (2012), used here as a comparison device rather than as a rationality test, in a line that runs back to Nordhaus (1987); Grant et al. (2026) develop a different extension of pairwise comparison to several horizons. Appendix B adds one small formal point — the cross-model revision contrast is free of the realised outcome, which is why it separates models on a sample whose loss comparisons resolve nothing. That flexible neural models often fail to beat simple benchmarks in macroeconomic nowcasting is close to a consensus, and our headline result agrees with it rather than overturning it. What we add is narrower: a published density neural MIDAS model, implemented to specification and confronted with a benchmark its original comparison set did not contain, together with a quantification of how little a comparison of

this size can settle — and a demonstration that the same sensitivity applies to our own presentation of our own estimator.

The paper proceeds as follows. Section 2 states the evaluation design, what was frozen in advance, and the comparison set together with the two neural estimators under evaluation. Section 3 reports the results, asks why the neural arms move so little (Section 3.5), and documents four numerical failures, one of which can silently disable a model confidence set (Section 3.6). Section 4 then develops the comparator argument, and Section 5 measures what the design could have detected. Appendices record the data construction, two short formal notes that the results motivated, and the model specifications and scoring conventions together with the sensitivity and accounting tables of the revision.

2 Evaluation design and estimators

The design is fixed in a protocol frozen before the scoring reported here, and the manuscript reports what that protocol produces. Section 2.6 gives the terms of the freeze; the remainder of this section states the design and then the comparison set it is applied to.

2.1 Target, origins and the horizon axis

The target is quarter-on-quarter growth in US real GDP at an annual rate, as first published: the advance release rather than a later vintage. Appendix A gives the construction. Forecasts are formed at *monthly* origins. Writing mtq for the number of months from the origin month to the final month of the target quarter,

$$mtq = 12\,y^{\text{target}} + 3\,q^{\text{target}} - \left(12\,y^{\text{origin}} + m^{\text{origin}}\right),$$

each target quarter is forecast at ten origins, from $mtq = 8$ down to $mtq = 0$ and then once more at $mtq = -1$, a backcast origin falling after the quarter has closed but before the advance release. Following the convention of Kim and Swanson (2018) these group into four blocks: two-quarter-ahead forecasts ($mtq \in \{8, 7, 6\}$), one-quarter-ahead forecasts ($\{5, 4, 3\}$), nowcasts ($\{2, 1, 0\}$) and the backcast.

The three nowcast origins are the headline set. They are the only region in which the monthly information set grows *inside* the target quarter, so they are the only place where the question of this paper — whether a model converts a month of new data into a better predictive distribution — has content. Origins at $mtq \geq 3$ and the backcast are reported but are not part of the pre-specified headline set.

Resolving the axis this way is not incidental to the results. Section 4.2 shows that pooling the four blocks into one number can reverse the ordering of two models on the same forecasts.

2.2 Information sets and the asymmetry with the benchmark

Two information sets are constructed. The headline set, *monthly vintage*, is release-aware: at each origin a model sees only the monthly series actually published by that date, at the vintage then available, so publication lags and revisions enter the design as they would in practice. The second, *full pseudo-real-time*, relaxes the vintage requirement while retaining the origin timing; it is a looser comparison and is reported as a robustness check.

The mixed-frequency models update at every monthly origin. The autoregressive benchmark cannot: it conditions on quarterly GDP alone and therefore updates only when a new advance release enters the information set, which happens once per quarter. Its forecast path inside a quarter is a step function. This asymmetry is deliberate and is the point of the comparison. A benchmark that cannot see the monthly flow is the natural floor for the question of whether the monthly flow is worth anything, and the appropriate reading is not that the comparison is unfair to the benchmark but that any model with access to more information should be expected to clear it.

2.3 Estimation splits

Models are fitted on release-aware nested splits at eleven expanding cutoffs. Within a cutoff the training data are split again, and every hyperparameter, stopping decision and model selection uses the inner validation fold only; the outer block is scored once and never consulted during fitting. Neural arms are fitted at five random seeds and their predictive draws pooled, so that a reported forecast is not a single lucky initialisation. The design is pseudo-real-time rather than real-time: the vintages are historical and the computation is retrospective.

2.4 Losses, weighting and inference

The primary loss is the continuous ranked probability score, computed from the pooled predictive draws by the off-diagonal U-statistic estimator, so that it is unbiased for the population CRPS at the draw count used. Squared error and absolute error are reported as secondary losses, and Section 3.4 shows the choice between them matters for what can be excluded. Throughout, the gain of a model m over a benchmark b is $100\,(\bar{L}_b - \bar{L}_m)/\bar{L}_b$, in per cent of the benchmark's loss and positive when the model is better; relative CRPS in the tables is $\bar{L}_m/\bar{L}_b$, below one when the model is better.

A target quarter appears at several origins, and treating each origin-quarter pair as an independent observation would let repeatedly-forecast quarters dominate. Losses are therefore averaged *within* a target quarter first, at a fixed *mtq*, and target quarters then weighted equally.

Inference is by a circular moving-block bootstrap over target-quarter clusters, with block length five quarters, 4,999 replications and a fixed seed. Where several origins form one family of comparisons, p-values are adjusted by Holm's method (Holm, 1979) within the family defined by an evaluated estimator, a baseline and an information set. Two boundaries of the inference are worth

Table 1: Evaluation sample. One row of the master panel is one model, target quarter and forecast origin; *mtq* is the number of months from the origin to the end of the target quarter.

Information set	Models	Target quarters	Pre-COVID	Origins	mtq range	Rows
full_pseudo_rt	7	89	64	264	-1 to 8	6,027
hvc_monthly_core4	8	89	64	264	-1 to 8	6,888
monthly_vintage	13	89	64	264	-1 to 8	11,193

Scored rows only. The IQ-HVC arm carries point forecasts and is scored on squared error.

stating plainly. The Holm families are the per-estimator families just defined; statements that range across models, regimes and losses — such as the sign counts of Section 3 — are descriptive tallies and carry no family-wide adjusted p-value. And the reported intervals are unadjusted percentile intervals from the circular block bootstrap (Politis and Romano, 1992); the Holm adjustment applies to p-values only. The primary pre-specified comparison is each evaluated estimator against its tier-2 comparators under CRPS at the three nowcast origins. Asymptotic Diebold–Mariano and Giacomini–White tests are not used as the primary inference; with 64 target quarters and the loss concentration documented in Section 5.2, their normal approximation is not credible here.

Results are reported for the pre-COVID sample (2004–2019, 64 target quarters) and for a COVID-excluded sample (2004–2025 less 2020–2022, 74 quarters). A full-sample view and the COVID quarters themselves are reported for contrast only: the 2020 quarters are extreme enough to reverse full-sample orderings by themselves, and no test is run on them.

Table 1 gives the resulting evaluation panel. One row is one model, target quarter and forecast origin.

2.5 A diagnostic for whether forecasts move

Because loss differences turn out not to separate most of these models, we also report a statistic that does not pass through the realised outcome and therefore has much smaller sampling variance. For a fixed target quarter and a step from an earlier origin to a later one — forward in calendar time, from larger *mtq* to smaller — define the revision and the prior-origin error

$$r = f(mtq_{\text{later}}) - f(mtq_{\text{earlier}}), \qquad e = y - f(mtq_{\text{earlier}}),$$

and report the revision size $\text{sd}(r)/\text{sd}(y)$ together with the slope β from regressing e on r. A β near one means the revision incorporates the news it should; a β near zero means the model does not revise toward the outcome. This is the multi-horizon logic of Patton and Timmermann (2012) used as a comparison device rather than as a rationality test of a single forecaster.

Two limitations belong with the statistic and not in a footnote. First, a large $|\beta|$ paired with a very small revision is noise amplification rather than efficiency, so the two numbers must be read together. Second, the statistic measures movement and not its direction, which bounds what it can be used for; Section 3.3 states that bound where the evidence for it appears. Appendix B formalises

why a statistic with this limitation is nonetheless worth having: being free of the realised outcome, it escapes the heavy-tailed term that widens every loss-based interval in this panel.

2.6 Pre-specification and what it does not cover

The protocol above — the horizon axis, the information sets, the losses and weighting, the bootstrap and its seed, the regime definitions, the diagnostic formulas, and a list of claims the paper is not permitted to make — was frozen before the scoring reported here, and one decision rule about estimator capacity was sealed with a numeric threshold before the variant it judges was fitted. The frozen protocol, the fitted artefacts and the code that turns one into the other are under content-hash control, with an audit that regenerates every exhibit and every quoted number and confirms that no recorded output has drifted.

Two qualifications. The benchmark tiers of Section 2.7 were extended once, after the supplementary arms had already been scored under their own sealed rule and in a direction that narrowed rather than widened this paper's claims; the amendment moved already-scored benchmarks between reporting tiers and altered no threshold. And the model confidence set, the power analysis of Section 5 and the fluctuation test were added after the results were seen. They are reported statistics, not decision gates: no verdict in this paper turns on them, and they are labelled as post-hoc wherever they appear. The two short formal notes of Appendices B and D were likewise written after — and because of — those results, and explain them rather than gate anything.

2.7 Three tiers of benchmark

Section 4 argues that the composition of a comparison set decides what a comparison can show. The set used here is therefore structured by the question each member answers rather than by family resemblance.

Tier 0 asks whether the target is predictable at all. It contains an expanding historical mean and a random walk on the last released growth rate, both updating at each GDP release. *Tier 1* asks whether the monthly information set has any value, and contains the Student-t autoregression described in Section 2.2, which updates quarterly and is a floor rather than a competitor. *Tier 2* asks the question that matters for a proposed estimator — whether it uses the monthly information set better than simple models seeing the same data — and contains models that update monthly: a Student-t dynamic factor model with a bridge equation, two classical MIDAS specifications (Student-t MIDAS and quantile MIDAS), two raw-lag neural arms, quantile gradient boosting on fixed Almon summaries of the monthly panel, and a combined bivariate ADL bridge averaging one regression per monthly series.

Confirmatory claims about a proposed estimator are made against tier 2. Tier 1 is reported throughout because it is the informative floor, and because clearing it turns out to be harder than the literature suggests.

2.8 The estimators under evaluation

Two neural density estimators are evaluated against that set.

The first is a structured CRPS-MIDAS estimator. It places a learned aggregation layer between the raw monthly lags and a neural head, so that the mapping from monthly data to a quarterly predictive distribution is factored through a low-dimensional summary rather than learned from lags directly. It is trained by CRPS on pooled predictive draws, with a spectral-norm constraint on the generator, a bounded output transform, and early stopping on inner validation CRPS. Two raw-lag variants remove the aggregation layer and are reported in tier 2: one matched to the structured arm's parameter count and one left unconstrained.

The second is QRNN-MIDAS (Xu et al., 2021), an external published model included so that the comparison is about a class of estimators rather than about our own implementation of one. We implement it to the authors' specification: a learned exponential-Almon alignment per series, one hidden layer with a tanh transfer, a linear output, and the six predictors their US GDP application names — industrial production, consumer and producer prices, M2, the oil price and the S&P 500. Because that predictor restriction is itself a modelling choice, we also fit the same architecture to all 126 series of the monthly panel. Section 3.2 shows the two versions falling on opposite sides of the model confidence set, which is consistent with specification restraint, rather than architecture, doing the work.

All arms are scored on the same origins, the same target quarters and the same losses, and every arm in tiers 0 to 2 sees the information set its tier permits.

3 Results

This section reports what the evaluation produces: accuracy by within-quarter origin, the set of models the data can exclude, and whether the forecasts move when a month of new information arrives. The headline sample throughout is the release-aware monthly-vintage information set on the 64 pre-COVID target quarters (2004–2019), scored separately at the three within-quarter origins $mtq \in \{2, 1, 0\}$, with the COVID-excluded sample (74 quarters) reported alongside. All intervals are 95% percentile intervals from the circular moving-block bootstrap over target-quarter clusters described in Section 2. Following the pre-specified reporting rule we give point estimates and intervals rather than binary verdicts.

3.1 Accuracy by within-quarter origin

Table 2 reports CRPS relative to the quarterly-updated AR benchmark, separately at each origin; Table 3 repeats the exercise under squared error. Figure 1 plots the same quantities as information accrues within the quarter.

Two patterns are stable and one apparent pattern is not.

The first stable pattern concerns the neural arms. At $mtq = 1$ the published QRNN-MIDAS specification of Xu et al. (2021) scores 1.103 times the AR benchmark's CRPS, the structured

Table 2: Pre-COVID CRPS relative to the quarterly-updated AR benchmark, by forecast origin (monthly-vintage information set). Values below one beat AR.

Model	8	7	6	5	4	3	2	1	0	-1	95% CI mtq=2	95% CI mtq=1	95% CI mtq=0
DFM-t	1.046	1.055	1.000	1.483	1.808	1.448	0.872	0.841	0.973	1.077	[-8, 37]	[-5, 42]	[-20, 25]
Quantile gradient boosting	1.007	1.000	0.970	0.951	0.947	0.919	0.941	0.925	1.003	1.016	[-8, 22]	[-5, 23]	[-18, 14]
Random walk	1.373	1.354	1.329	1.223	1.222	1.213	1.196	0.987	1.080	1.093	[-43, -1]	[-18, 29]	[-22, 9]
Historical mean	1.254	1.251	1.224	1.137	1.141	1.129	1.081	1.075	1.185	1.147	[-30, 10]	[-29, 11]	[-40, 0]
Structured CRPS-MIDAS	1.172	1.078	1.035	1.015	1.000	1.004	1.050	1.037	1.142	1.217	[-31, 14]	[-33, 15]	[-64, 17]
QRNN-MIDAS (published)	1.359	1.316	1.281	1.288	1.296	1.230	1.130	1.103	1.215	1.108	[-34, 9]	[-28, 10]	[-37, -9]
Combined bivariate ADL bridge	0.920	0.941	0.909	1.209	1.354	4.006	1.521	1.184	2.517	12.892	[-151, 15]	[-75, 24]	[-407, 13]
Raw-lag (capacity matched)	1.707	1.712	1.569	1.599	1.521	1.518	1.595	1.532	1.702	1.729	[-127, -11]	[-122, -7]	[-149, -17]
Raw-lag (unmatched)	2.555	2.589	2.674	2.904	2.798	2.939	2.893	2.769	3.015	3.052	[-335, -81]	[-307, -74]	[-355, -78]

Target-quarter-equal CRPS; intervals are quarter-cluster moving-block bootstrap gains in % of AR loss. mtq=0 is an origin in the quarter's last month; mtq=-1 is the backcast origin.

Table 3: Root mean squared forecast error relative to the quarterly-updated AR benchmark, by forecast origin (pre-COVID, monthly-vintage). The point-forecast counterpart of Table 2, reported in the loss and horizon layout of Kim and Swanson (2018).

Model	8	7	6	5	4	3	2	1	0	-1
DFM-t	1.046	1.012	0.995	0.990	1.057	0.989	0.883	0.846	1.030	1.091
Quantile gradient boosting	1.070	1.056	1.008	1.015	1.012	0.994	0.979	0.967	1.132	1.082
Random walk	1.403	1.397	1.361	1.322	1.222	1.232	1.212	0.924	1.083	1.087
Historical mean	1.213	1.205	1.174	1.093	1.089	1.089	1.040	1.036	1.222	1.160
Structured CRPS-MIDAS	1.229	1.115	1.060	1.120	1.107	1.138	1.207	1.217	1.471	1.506
QRNN-MIDAS (published)	1.299	1.258	1.213	1.262	1.264	1.227	1.020	1.018	1.195	1.110
Combined bivariate ADL bridge	1.007	1.005	0.961	1.312	1.420	6.740	1.688	1.185	3.614	23.502
Raw-lag (capacity matched)	1.728	1.709	1.676	1.675	1.620	1.664	1.799	1.771	2.120	2.145
Raw-lag (unmatched)	2.905	2.831	3.022	3.179	2.916	3.099	2.623	2.497	3.052	3.137

Target-quarter-equal squared error. The horizon index of Kim and Swanson ($h = -1$ backcast, $1, 2, 3$ nowcast, $4, 5, 6$ one quarter ahead, $7, 8, 9$ two quarters ahead) maps to mtq as $7, 8, 9o8, 7, 6$; $4, 5, 6o5, 4, 3$; $1, 2, 3o2, 1, 0$.

estimator of this paper scores 1.037, and the two raw-lag arms score 1.532 and 2.769. Widening to the pre-specified headline cells — two regimes × three origins × three losses, eighteen in all — each of the four neural arms performs worse than the AR benchmark in *all eighteen*, with deficits for QRNN-MIDAS ranging from 3.7 to 43.5 per cent. For the capacity-matched raw-lag arm and for QRNN-MIDAS fitted to all 126 series the interval excludes zero in all eighteen cells; for the author-specified QRNN-MIDAS it does so in six, and for the structured estimator in three. The sign is thus consistent across every design choice we varied, even where the intervals are wide.

The second stable pattern is that this does not convert into a positive finding for anything else. The two arms with the best point estimates are the Student-t dynamic factor model (0.841 at $mtq = 1$) and quantile gradient boosting (0.925), and neither can be said to beat the AR benchmark. Across the same eighteen cells the interval excludes zero in *none*, and the point estimate itself changes sign in nine: the factor model's gain ranges over $[-18.4, +28.4]$ per cent and the boosted quantile model's over $[-28.1, +7.5]$. Section 5 quantifies why. At 64 target quarters the interval half-width is roughly twenty percentage points of the benchmark's loss, so a 15.9 per cent point estimate is not a detection. We therefore report these orderings as point estimates and claim no outperformance for either model.

The apparent pattern that does not survive is the ordering within the middle of the field. The random walk, the historical mean, the combined bivariate ADL bridge and the author-specified QRNN-MIDAS occupy a band between 0.99 and 1.22 at $mtq = 1$ whose members reorder across regimes, losses and origins.

Two features of Table 2 should be separated from the substantive comparison. First, the classical MIDAS benchmarks take values orders of magnitude above the rest, and the ADL bridge at the backcast origin more than ten times the benchmark. These are numerical failures of our implementations under this protocol rather than forecasting results, and Section 3.6 treats them as such; they are excluded from the headline comparisons by the pre-specified filter. Second, the backcast column ($mtq = -1$) behaves differently from the three within-quarter origins because the target quarter has already closed. We report it for completeness and exclude it from the pre-specified headline set.

3.2 What the data can exclude

Pairwise comparisons against a nominated benchmark answer whether one model beats another. They do not answer which models the data can rule out at all, and running many of them invites a multiplicity objection that Holm correction within a single family does not fully address. Table 4 therefore reports the model confidence set of Hansen et al. (2011), constructed by the range procedure on the same quarter-cluster bootstrap used elsewhere in the paper. It nominates no benchmark a priori and carries the multiplicity internally.

The central result is what the set retains. Under the density score at the 90 per cent level it excludes QRNN-MIDAS fitted to all 126 series and the unmatched raw-lag arm in all six cells, and the capacity-matched raw-lag arm at the 75 per cent level. It excludes nothing else. In particular,

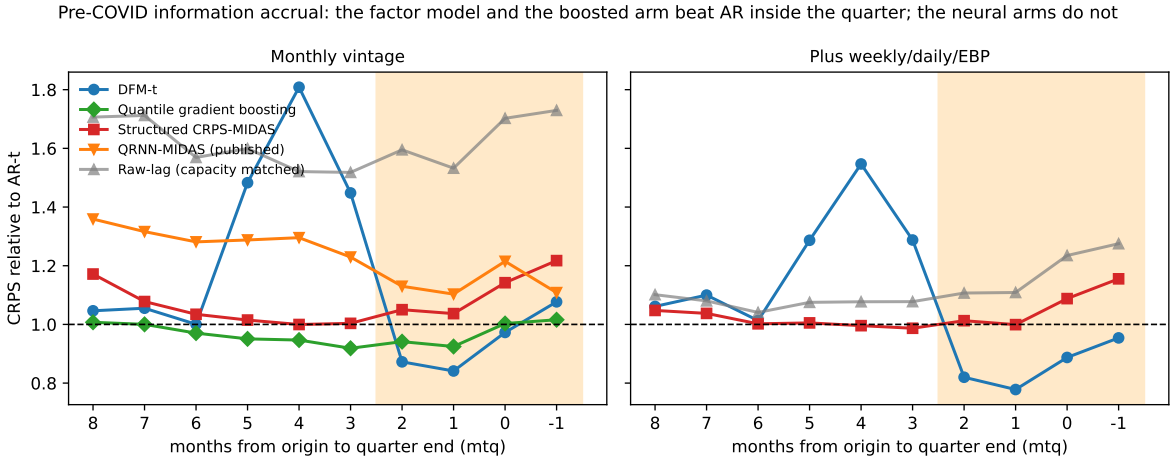

Figure 1: CRPS relative to the quarterly-updated AR benchmark as information accrues within the target quarter, pre-COVID. Values below one favour the model.

Table 4: Model confidence set. Membership across the six cells formed by two regimes (pre-COVID, COVID-excluded) and the three within-quarter origins, for the monthly-vintage information set.

Model	Rel. CRPS	CRPS 90%	CRPS 75%	MSFE 90%	MSFE 75%
DFM-t	0.841	✓	✓	✓	✓
Quantile gradient boosting	0.925	✓	✓	✓	✓
Random walk	0.987	✓	✓	✓	✓
AR-t (quarterly update)	1.000	✓	✓	✓	✓
Structured CRPS-MIDAS	1.037	✓	✓	✓	✓
Historical mean	1.075	✓	✓	✓	✓
QRNN-MIDAS (published)	1.103	✓	✓	✓	✓
Combined bivariate ADL bridge	1.184	✓	✓	✓	✓
Raw-lag (capacity matched)	1.532	5/6	0/6	✓	5/6
QRNN-MIDAS, all 126 series	1.636	0/6	0/6	✓	2/6
Raw-lag (unmatched)	2.769	0/6	0/6	✓	0/6

✓ marks membership in all six cells; a fraction gives the number of cells retained. Relative CRPS is the pre-COVID mean at mtq = 1 divided by that of the quarterly-updated AR. The set is constructed by the range procedure of Hansen, Lunde and Nason (2011) on the same quarter-cluster circular moving-block bootstrap used everywhere else in the paper (block 5, 4,999 replications), after excluding the two arms that explode numerically. Neither the historical mean nor the random walk is excluded in any cell, which bounds what this sample can establish; and no arm at all is excluded by the 90 per cent set under squared error.

neither the historical mean nor the random walk is excluded in any cell: 72 rows for each of the two, spanning two regimes, three origins, two losses, two confidence levels and three treatments of the numerically failed arms. The eight models that survive even the 75 per cent set — among them the AR benchmark, the factor model, the boosted quantile model, the author-specified QRNN-MIDAS and the structured estimator of this paper — cannot be separated by this procedure on this sample; that is a statement about the procedure's resolution here, not a claim that the models are equivalent.

This bounds what can be claimed in either direction. A size study on null data (400 replications per configuration) puts the procedure's empirical rejection rate at 0.117 for thirteen exchangeable models and 0.140 for four, against a nominal 0.10: the procedure is mildly size-inflated, eliminating somewhat too readily, which makes its failure to eliminate any of these arms more informative rather than less. Reporting the set at both the 90 and 75 per cent levels follows the practice of Hansen et al. (2011).

The models the set does exclude share a feature. All three are neural arms in unrestricted specifications: the two raw-lag arms of this study, and QRNN-MIDAS fitted to the full FRED-MD panel rather than to the six series that Xu et al. (2021) specify. The author-specified QRNN-MIDAS, which differs from the excluded variant only in that restraint, survives in every cell, as does the structured estimator of this paper. Membership therefore tracks specification restraint rather than architecture in this set — though with a single pair carrying the contrast, this is consistency with that reading, not proof of it.

3.3 Do the forecasts move?

Loss differences do not separate most of these models. A diagnostic that does not pass through the realised outcome has far smaller sampling variance, and Table 5 reports one: the size of the revision a model makes when a month of new data arrives, scaled by the standard deviation of the target, in the spirit of the multi-horizon restrictions of Patton and Timmermann (2012).

Between the first and second within-quarter origins the factor model revises by 0.398 standard deviations of the target and the author-specified QRNN-MIDAS by 0.268, while the AR benchmark — which cannot see the monthly data at all and moves only when a new quarterly release arrives — revises by 0.049 and the historical mean by 0.044. The structured estimator revises by 0.089 and quantile gradient boosting by 0.093.

The scope of this diagnostic must be stated with it. Revision size measures how much a model moves, not whether it moves to the right place. The pair `gbm_quantile` and `structured_crps` makes the point: they revise by essentially the same amount (0.093 against 0.089) and differ by eleven percentage points in relative CRPS. The diagnostic separates models that use the monthly flow from models that cannot, in a place where the loss comparison does not; it does not rank models that move similarly. Appendix B gives the reason for the asymmetry: the contrast never touches the realised outcome, so its interval is free of the outcome's tails, and on this pair it is six to seven times as precise as the loss comparison on the same quarters.

Table 5: Do the forecasts move when a new month of data arrives? Pre-COVID, forward in time.

Model	Information set	Revision mtq 2→1	Revision mtq 1→0	Slope β
AR-t (quarterly update)	monthly_vintage	0.049	0.480	1.20
Combined bivariate ADL bridge	monthly_vintage	1.662	2.919	0.22
DFM-t	monthly_vintage	0.398	0.415	0.55
Quantile gradient boosting	monthly_vintage	0.093	0.090	2.97
Historical mean	monthly_vintage	0.044	0.041	2.41
MIDAS-t	monthly_vintage	180858.823	152345.919	0.68
QRNN-MIDAS (published)	monthly_vintage	0.268	0.251	0.71
QRNN-MIDAS, all 126 series	monthly_vintage	0.198	0.366	0.62
Quantile MIDAS	monthly_vintage	387.182	389.834	3.64
Random walk	monthly_vintage	1.020	0.043	0.85
Raw-lag (capacity matched)	monthly_vintage	0.165	0.213	0.83
Raw-lag (unmatched)	monthly_vintage	0.360	0.382	1.66
Structured CRPS-MIDAS	monthly_vintage	0.089	0.096	-0.61
AR-t (quarterly update)	full_pseudo_rt	0.053	0.485	1.15
DFM-t	full_pseudo_rt	0.368	0.397	0.72
MIDAS-t	full_pseudo_rt	82445.854	159336.623	0.12
Quantile MIDAS	full_pseudo_rt	106.860	97.252	2.93
Raw-lag (capacity matched)	full_pseudo_rt	0.120	0.133	1.31
Raw-lag (unmatched)	full_pseudo_rt	0.499	0.478	0.41
Structured CRPS-MIDAS	full_pseudo_rt	0.058	0.079	-1.50
Core4 factor bridge	hvc_monthly_core4	0.304	0.457	1.27
FRED-MD factor MIDAS	hvc_monthly_core4	0.341	0.416	1.16
Horizon-specific surface	hvc_monthly_core4	0.026	0.226	1.74
Almon MIDAS ridge	hvc_monthly_core4	0.286	0.515	1.25
Shared HVC surface	hvc_monthly_core4	0.036	0.226	1.69

Revision size is $\mathrm{sd}(f_{\text{later}} - f_{\text{earlier}})/\mathrm{sd}(y)$. β regresses the prior-origin error on the cumulative revision: $\beta \approx 1$ means the revision incorporates the news, $\beta \leq 0$ means it moves away from the outcome. A large $|\beta|$ with a tiny revision is noise amplification, not efficiency. The IQ-HVC panel lists only the arms in its declared comparison set.

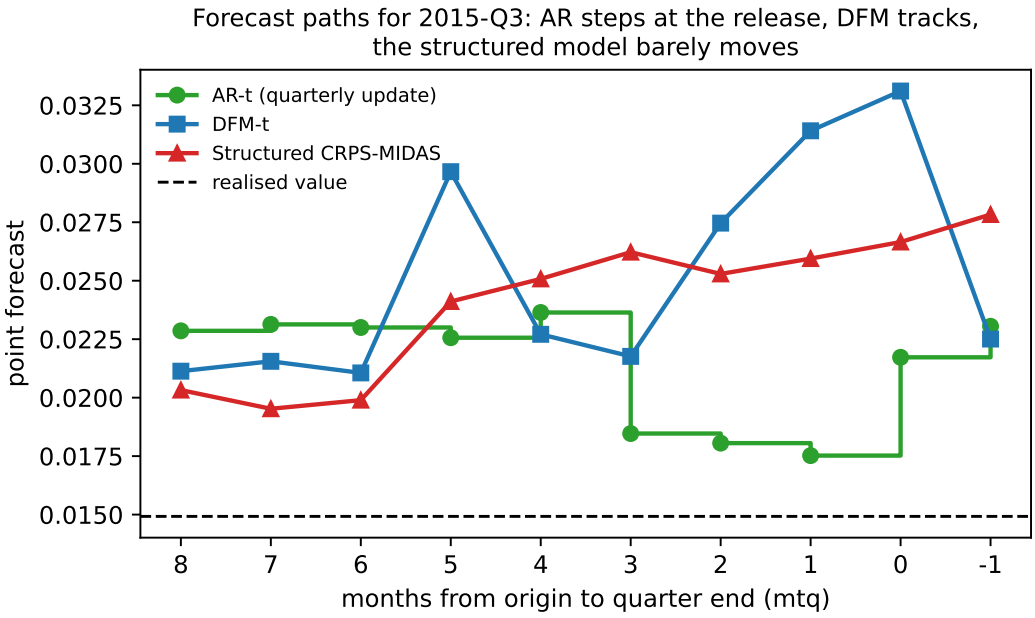

Figure 2: Forecast paths for a single target quarter, 2015:Q3, as the origin advances through the quarter. The quarter is a mid-expansion illustration fixed before the statistics were computed; the sample-wide evidence is Table 5, not this picture.

Figure 2 shows what these numbers look like for a single target quarter. The factor model's path moves visibly as each month arrives; the AR benchmark's is a step function by construction; the structured estimator's is nearly flat.

Within the movement itself, one arm behaves differently from the rest. Measure a revision's magnitude by its root mean square; that magnitude decomposes into a constant drift and a news-driven dispersion, the standard deviation reported in the table being the dispersion alone, and the two coincide only when the mean revision is zero. By that decomposition the structured estimator carries 28.5 per cent of its squared revision magnitude as a fixed shift at $mtq\colon 2 \to 1$, and between 8.7 and 35.9 per cent across the six regime-by-step combinations. Every other arm outside the raw-lag family stays below nine per cent and mostly below one; the highest is the AR benchmark itself at 8.1, on a revision so small that a negligible mean is a large fraction of it. The structured estimator does not merely move less than the factor model; the part of its movement that does exist is largely not a response to news.

3.4 Density versus point scores

The exclusions in Section 3.2 occur under the density score only. Under squared error at the 90 per cent level the model confidence set retains all eleven arms in every cell; at the 75 per cent level it excludes the unmatched raw-lag arm everywhere and the 126-series QRNN-MIDAS variant in four of six cells. The ordering in Table 3 is broadly that of Table 2, but the separation is smaller.

The reading we can defend is narrower than the one that first suggests itself. Exclusion evidence

Table 6: Why the structured estimator does not respond to monthly news. Each row changes one thing relative to the estimator of Section 3 and is scored on the same sub-sample.

Variant	Sub-sample	Revision mtq 2→1	CRPS gain vs AR (%)	95% CI
trained on mtq=2 origins only	11 cutoffs, 5 seeds	0.082	-0.8	[-23, 19]
trained on within-quarter origins only	11 cutoffs, 5 seeds	0.159	-11.6	[-49, 17]
Structured CRPS-MIDAS	11 cutoffs, 5 seeds	0.089	-3.7	[-30, 15]
no early stopping (full 800-step budget)	5 cutoffs, 3 seeds	0.288	-154.4	[-243, -81]
no spectral norm / no bounded output	5 cutoffs, 3 seeds	0.112	-10.8	[-31, 8]
both of the above	5 cutoffs, 3 seeds	0.448	-277.6	[-567, -84]
reduced factor dimension, lower lr, more draws	5 cutoffs, 3 seeds	0.102	-32.7	[-69, -4]
Structured CRPS-MIDAS	5 cutoffs, 3 seeds	0.118	-10.9	[-24, 4]

Pre-COVID, mtq=1, monthly-vintage. Positive gain means better than AR. The reduced-capacity row changes the factor dimension, learning rate, draw count and early-stopping patience together; it lowers the parameter count by only 1.3%, so it does not settle whether capacity or structure is responsible.

is observed under the density score and not, at the same level, under squared error; whether that reflects the predictive distributions rather than the two statistics' different power on this panel, the sample cannot say. It is in any case a limitation on the claims: every statement in this paper about what the data can exclude is a statement under CRPS, and the corresponding statement under squared error is weaker.

3.5 Why the neural arms do not respond

Section 3.3 establishes that the structured estimator revises by 0.089 standard deviations of the target between the first two nowcast origins, against 0.398 for the factor model, and that a quarter or more of even that movement is a fixed drift rather than a response to news. This subsection asks why, by varying one element of the fitting procedure at a time and reporting what each purchases. The exercise is diagnostic and internal to the model class; it is not a claim that the answer generalises beyond it.

Table 6 reports revision size and CRPS against the AR benchmark for each variant. The two training-origin variants run on the full eleven cutoffs, where the headline arm's revision is the 0.089 above; the remaining variants run on a five-cutoff sub-sample, where the headline arm's revision is 0.118, and each contrast below is made within its own sample. The clearest pattern is that the levers which increase movement also destroy accuracy.

Removing early stopping and training for the full budget raises the revision from 0.118 to 0.288, more than doubling it, while the CRPS deficit against the benchmark widens from 10.9 to 154.4 per cent with an interval far from zero. Removing the spectral-norm constraint and the bounded output together with early stopping raises the revision to 0.448 and the deficit to 277.6 per cent. Removing the constraints alone is milder in both directions. The restraints that make the estimator inert are the same restraints that keep it from diverging: what is on offer is a frontier, not a fix.

Restricting the training origins does not help either. Fitting only on the $mtq = 2$ origins gives a revision of 0.082, slightly *below* the headline arm's 0.089; fitting on all three within-quarter origins gives 0.159 with a wider deficit. Neither interval separates from the headline arm.

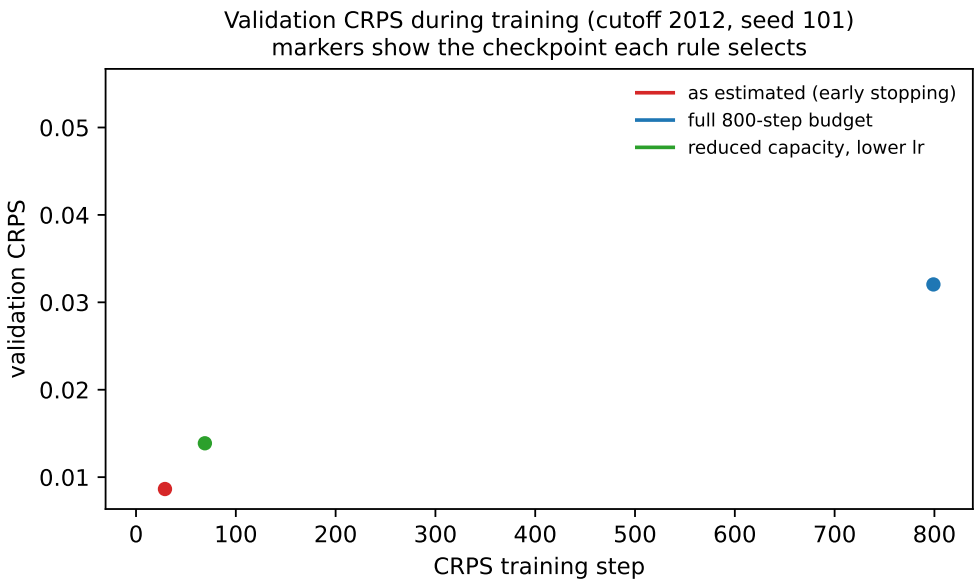

Figure 3: Inner-validation CRPS during pre-training of the structured estimator, for one cell (cutoff 2012, seed 101), shown as an illustration. That the early minimum is systematic rather than particular to this cell is established by the no-early-stopping ablation of Table 6.

Figure 3 shows the mechanism behind the first result. Validation CRPS is minimised very early in training, so early stopping selects a checkpoint close to initialisation — a model that has learned little and therefore moves little. Training past that point buys movement and loses accuracy, exactly as the table reports.

A natural reading of the above is that the estimator is simply too small. It is not. A pre-specified sweep over the width of the aggregation block, from 3,575 to 50,855 generator parameters, moves neither responsiveness nor accuracy systematically. Revision size at $mtq\colon 2 \to 1$ runs 0.159, 0.145, 0.240, 0.156 across widths 8, 16, 64 and 128 — not monotone — and the widest-minus-narrowest difference is -0.003 with a 95% interval of $[-0.035, +0.028]$. On accuracy the smallest model (-9.5 per cent against the benchmark) is no worse than the incumbent (-14.8) and the largest is worse than both (-33.6). The sealed rule attached to this sweep required a revision of at least 0.25 and a positive gain to declare capacity the binding constraint; the maximum revision observed was 0.240 and every gain was negative, so the rule fails on both counts. A separately pre-specified low-capacity variant fails the same rule from the other direction, with a revision of 0.102 and a deficit of 32.7 per cent.

The contrast with the raw-lag arms is the substantive result. There, where no aggregation layer stands between the monthly lags and the neural head, raising capacity does buy movement: the same comparison gives a revision difference of $+0.201$ with an interval of $[+0.067, +0.310]$ that separates from zero. In the structured family a fourteen-fold increase in parameters buys nothing measurable.

We state the reading narrowly. The aggregation layer appears to make responsiveness insensitive

to capacity, which is not the same as saying it causes the failure. The comparison that would settle the matter is unavailable here: quantile gradient boosting, which does clear the benchmark on point estimates while revising by a similar 0.093, differs from the structured arm in three ways at once — a fixed rather than learned aggregation, a tree ensemble rather than a neural head, and a quantile rather than a CRPS objective. This sample cannot separate the three, and we leave the question open in Section 6 rather than resolve it by assertion.

3.6 Numerical robustness

Four numerical failures occurred in the course of this study, all of them in classical or external comparators rather than in the structured estimator, whose bounded output makes divergence impossible by construction. We report them because two of them would silently change the conclusions of a reader who did not look.

On the monthly-vintage panel, at the three nowcast origins that carry the headline comparisons, the Student-t MIDAS arm reaches CRPS of order 10^5 times the AR benchmark and the quantile MIDAS arm between 382 and 454 times. Both are better behaved at the longest horizons and worse in the middle of the axis, which is itself a sign that the failure is numerical rather than a matter of forecast difficulty. These are our implementations under a release-aware protocol with no regularisation, not a claim about the estimators as their authors specify them. A pre-specified filter flags any row whose predicted mean exceeds a threshold the target scale cannot reach, and the headline comparisons are computed with those rows excluded.

The combined bivariate ADL bridge averages one regression per monthly series. At one of eleven cutoffs, two of 119 usable series produce near-collinear regressions whose predictions dominate the equal-weighted average, lifting the mean forecast to 0.310 against an actual standard deviation of 0.021 while the median forecast stays at 0.027. At the other cutoffs mean and median nearly coincide. We excluded series whose inner-training aggregate has zero variance, since regression on a constant is undefined, but did not raise the threshold further: an arbitrary cutoff chosen after seeing the result would be tuning. The consequence is that this arm's poor showing should not be read as evidence about bridge equations in general. It was included as insurance for the argument that a linear factor model converts monthly data into accuracy, and it failed to provide that insurance; we report that rather than drop the arm.

The threshold above is set to the target's scale and the ADL bridge's largest prediction falls just below it. A filter calibrated to one failure mode does not detect another, and we record the gap rather than retune the threshold to the case that escaped it.

The fourth failure is the one with consequences for inference. Run at the conventional 90% level on the pre-COVID sample with the diverging arms left in, the model confidence set excludes *nothing*: all thirteen arms survive, including the one whose loss is five orders of magnitude above the rest. The outliers that inflate the mean loss inflate the bootstrap standard error alongside it, and the studentised statistic returns to an unremarkable value. With those arms removed the same procedure excludes two. The masking is partial — at the 75% level the diverging arms

are eliminated on their own — so the accurate statement is not that the procedure is helpless against outliers but that at the conventional level a heavy-tailed loss can suppress every rejection. Run without a filter, the finite-sample procedure loses its ability to discriminate — a property of the procedure under extreme losses, not a statement about the forecasts. Appendix D records the mechanism and why, at this sample size and outlier configuration, no amount of additional divergence or bootstrap depth changes it.

4 Comparator choice decides the answer

The results of Section 3 are not the first evaluation of a density neural MIDAS model on US GDP, and they disagree with the published ones. This section argues that the disagreement is not about the models. Four choices, each defensible and each made before any model is fitted, move what the comparison reports — the ordering, the set of exclusions, or both: which models enter the comparison set, how finely the horizon axis is resolved, how large the comparison set is, and which loss scores it. We take them in turn, and we make no claim that the published comparisons are wrong on their own terms.

4.1 Which models enter the set

Xu et al. (2021) evaluate QRNN-MIDAS on quarterly US GDP growth against QR-MIDAS and QRNN, reporting gains on test data of roughly 15 to 47 per cent against QR-MIDAS and 8 to 38 per cent against QRNN across five quantile levels.[1] The comparison is careful: it uses a rolling-window design, reports dispersion across windows, and applies a Diebold and Mariano (1995) test to each pair.

The feature we draw attention to is the composition of the comparison set rather than the conduct of the comparison. In the tables reporting that experiment the columns are QR-MIDAS, QRNN and QRNN-MIDAS. There is no autoregression, no random walk, no factor model and no unconditional mean — no benchmark, that is, from outside the neural and MIDAS families whose combination the paper proposes. Every reported gain is a gain over a member of the same family.

Implementing that model to its authors' specification and scoring it against a quarterly-updated AR benchmark at each within-quarter origin gives the result of Section 3.1: worse than the benchmark in all eighteen pre-specified headline cells, by 3.7 to 43.5 per cent. The two statements are compatible. A model can improve on every member of its own family and still not reach a benchmark that family does not contain.

We stop short of a stronger claim that the arithmetic in this paper does not support. Our own implementations of the classical MIDAS benchmarks are numerically fragile under this protocol (Section 3.6), and the margins by which the neural arms beat them here are correspondingly

[1] Read from their Tables 7 to 11, panel (b), over $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and the three lag configurations. Two entries report a QRNN-MIDAS loss an order of magnitude below every neighbouring value in the same table and appear to be typographical; the quoted range excludes them, and including them would widen its upper end.

Table 7: What a horizon-pooled, full-sample table reports, and what the same forecasts show once the sample and the origin are separated. Structured CRPS-MIDAS relative to AR-t.

Information set	Sample	Pooled over all origins	Nowcast window only
monthly_vintage	overall	0.813	0.829
monthly_vintage	pre_covid	1.066	1.074
monthly_vintage	covid_excluded	1.112	1.156
monthly_vintage	covid	0.646	0.647
monthly_vintage	post_2023	1.507	1.820
full_pseudo_rt	overall	0.777	0.780
full_pseudo_rt	pre_covid	1.026	1.032
full_pseudo_rt	covid_excluded	1.056	1.071
full_pseudo_rt	covid	0.621	0.617
full_pseudo_rt	post_2023	1.311	1.389

Values below one favour the structured model. Its full-sample advantage is produced entirely by the 2020–2022 quarters.

enormous and uninformative. Those margins are a fact about our implementations, not about the published ones, and we do not read them as evidence about the benchmarks used in the literature. The argument of this section rests on the composition of the comparison set, which is visible in the published tables, and on the addition of a benchmark from outside it.

4.2 How finely the horizon axis is resolved

A second choice is whether to report one number per model or one number per model and origin. Table 7 contrasts the two for the structured estimator of this paper.

Against the AR benchmark the two views agree closely. Against the factor model they do not. Pooling all origins, the structured estimator scores 0.921 times the factor model's pre-COVID CRPS and would be reported as the better of the two; resolving to the three within-quarter origins, it scores 1.203 and is the worse. The reversal holds in all four combinations of the two information sets and the two headline regimes, with pooled ratios of 0.921 and 0.947 on the monthly-vintage panel and 0.957 in both regimes on the full pseudo-real-time panel, against resolved ratios of 1.203, 1.202, 1.248 and 1.155.

The mechanism is not subtle. Pooling averages the within-quarter origins together with the backcast and the two forecast quarters, where an arm that is weak at nowcasting can be strong, and where the AR benchmark's inability to see monthly data costs it nothing. The horizon axis we use follows Kim and Swanson (2018); the point here is only that a horizon-pooled table and an origin-resolved table can rank the same two models in opposite orders on the same forecasts.

4.3 How large the comparison set is

A third choice is how many models to compare at once, and it interacts with any procedure that controls for multiplicity. The concern is an old one: White (2000) built the reality check precisely

Table 8: Sequential comparator addition on one fixed loss matrix: pre-COVID CRPS, monthly-vintage information set, $mtq = 1$, 64 target quarters, the two diverging classical arms excluded throughout.

Comparison set	Arms	Rank	Rel. CRPS	In 90% set	Excluded at 90%
QRNN-MIDAS family only	2	1	0.675	✓	QRNN-MIDAS, all 126 series
+ AR benchmark	3	2	1.103	✓	QRNN-MIDAS, all 126 series
+ historical mean, random walk	5	4	1.118	✓	QRNN-MIDAS, all 126 series
+ DFM-t	6	5	1.311	✓	QRNN-MIDAS, all 126 series
+ boosted quantile, ADL bridge	8	6	1.311	✓	QRNN-MIDAS, all 126 series
+ this paper's arms (all eleven)	11	7	1.311	✓	QRNN-MIDAS, all 126 series; Raw-lag (unmatched)

Rank, relative CRPS (against the best other arm in the set) and membership refer to the author-specified QRNN-MIDAS as the set grows; only the roster varies – panel, quarters, loss and bootstrap are fixed. Within its own family the model looks decisive; each added tier moves it down the ordering; no set can exclude it.

because searching many models against a benchmark overstates the best performer, Hansen (2005) sharpened its power, and the model confidence set descends from that line. The converse is less often remarked and is visible here. Because the model confidence set takes a maximum over the models in play, adding arms raises the critical value and shrinks what can be excluded.

Whether the effect bites here is a question an earlier draft of this paper answered with a contaminated example, drawing the contrast between a five-arm set on one information set and an eleven-arm set on another; the five-arm exclusion of the AR benchmark in that contrast turns out to be attributable to the information set, not to the roster, and we retract that attribution. The controlled comparison holds one loss matrix fixed — same panel, same quarters, same bootstrap — and varies only the roster. On that fixed matrix the AR benchmark is retained at every origin by the 75 per cent set whether five arms or eleven are in play. The roster effect that does survive the controlled comparison sits one arm further down and is modest: at the first two nowcast origins of the pre-COVID sample the five-arm roster excludes the capacity-matched raw-lag arm at the 90 per cent level — marginally at one of them, with a p-value of 0.0996 — and enlarging the roster to eleven rescues it at every origin, the larger maximum pushing the critical value past that arm's statistic.

Sequential addition on the same fixed matrix tells the story from the other end (Table 8). Starting from the two QRNN-MIDAS specifications alone, the family comparison looks decided: the author-specified version beats its sibling by a third and even the two-arm confidence set excludes the sibling. Adding the AR benchmark moves the author-specified version to second; adding the remaining tiers moves it to seventh of eleven; and at no step can any set exclude it. The model's reported standing is a creature of the roster, while the exclusions barely move — which is the comparator argument and the resolution argument of this paper in a single table.

This cuts against the presentation of our own results as much as anyone's, and it is why Section 3.2 reports the full set as the headline. It also means that the size of a model confidence set is not by itself a measure of evidence: our 90% set retains nine of eleven arms and the 75% set eight, and neither number would be comparable to a set built over a different roster.

4.4 Which loss scores the comparison

The fourth choice is the loss. As Section 3.4 reports, the exclusions in this paper occur under CRPS and not under squared error, where the 90% set retains all eleven arms in every cell. A study of the same models on the same data that scored conditional means rather than predictive distributions would find nothing to exclude at conventional levels.

4.5 What follows

None of the four choices is a mistake, and we have made each of them ourselves in one direction or another. The claim is narrower and, we think, harder to dismiss: on a sample of this size the ordering of density mixed-frequency nowcasts is not a property of the models alone. It is a joint property of the models and of a comparison design whose defensible variants disagree. Section 5 quantifies why the design has so little to say on its own — the differences at stake are smaller than the resolution of a 64-quarter comparison — and that is the same fact seen from the other side.

5 What this sample can and cannot establish

Section 3 reports a null: the model confidence set excludes neither the historical mean nor the random walk, and no model can be said to beat the AR benchmark. A null is informative only once paired with what the design could have detected, so this section measures that directly. It reports the difference the design resolves, how few target quarters supply that difference, what happens when they are removed, and whether a proper test finds the instability that removal suggests. None of this requires refitting; all of it runs on the scored panel.

5.1 The difference the design resolves

The half-width of a 95% interval is, in units of the benchmark's loss, the gain a model must post before its interval clears zero. The first two numeric columns of Table 9 report both quantities at $mtq = 1$ on the pre-COVID sample.

For the six arms whose CRPS lies nearest the benchmark's the half-width sits between 14.5 and 23.6 percentage points; the ADL bridge, whose implementation is numerically fragile (Section 3.6), is wider at 49.1. At 64 target quarters a model must improve on the AR benchmark's CRPS by roughly twenty per cent before this design can register the improvement, and the best model in the entire set achieves 15.9 per cent. That single comparison is the quantitative content of the null in Section 3.2: the exercise is not silent because the models are indistinguishable in truth, but because the resolution of a 64-quarter density comparison is coarser than the differences at stake.

The intervals themselves are, if anything, too short. A Monte Carlo calibrated to this design — 64 quarters, the same circular block bootstrap — puts the coverage of the nominal 95% percentile interval at 0.927 under an independent Gaussian differential, between 0.912 and 0.916 under

Table 9: What the design can detect. CRPS against the quarterly-updated AR at mtq = 1, pre-COVID, on 64 target quarters.

Model	Gain (%)	CI half-width (%)	Kurtosis	Top-3 share (%)	The three quarters
DFM-t	15.9	23.6	16.7	74	2009-Q1; 2008-Q4; 2009-Q4
Quantile gradient boosting	7.5	14.5	11.4	71	2009-Q3; 2009-Q4; 2012-Q4
Random walk	1.3	23.3	16.7	72	2009-Q1; 2008-Q4; 2009-Q2
Structured CRPS-MIDAS	-3.7	23.4	19.9	80	2009-Q2; 2009-Q1; 2009-Q4
Historical mean	-7.5	20.1	5.7	51	2009-Q3; 2004-Q1; 2005-Q3
QRNN-MIDAS (published)	-10.3	18.6	4.6	46	2009-Q4; 2005-Q2; 2006-Q2
Combined bivariate ADL bridge	-18.4	49.1	6.9	56	2018-Q2; 2019-Q3; 2018-Q3
Raw-lag (capacity matched)	-53.2	56.2	11.5	74	2018-Q4; 2018-Q2; 2019-Q2
QRNN-MIDAS, all 126 series	-63.6	46.5	1.3	35	2013-Q4; 2019-Q1; 2018-Q1
Raw-lag (unmatched)	-176.9	117.8	6.3	49	2018-Q4; 2018-Q3; 2019-Q2

Gain and half-width are per cent of the AR benchmark's loss; a model must post a gain exceeding the half-width before its interval clears zero, and none does. Kurtosis is of the per-quarter loss difference against AR, and the top-3 share is the fraction of its sum of squares contributed by the three largest quarters, which are named in the final column. For every arm that competes with AR those three quarters are the financial crisis; for the arms that lose badly they are that arm's own instability. This is why the interval does not narrow with sample length over the observed range (Figure 4).

Student-t tails down to 2.5 degrees of freedom, and between 0.872 and 0.923 when the 64 quarters are resampled from the observed loss pairs themselves, concentration included (Table 16 in Appendix C). Two readings follow. The shortfall is a small-sample property of the percentile interval under blocking, not of the heavy tails specifically: the Gaussian control loses almost as much coverage as the t with 2.5 degrees of freedom. And its direction strengthens the null rather than undermining it, since correctly calibrated intervals would be wider, and an interval that contains zero at nominal 95% keeps containing it when widened.

Whether lengthening the sample would help is a question the panel can answer within its own range. Drawing contiguous windows of T target quarters — contiguous, because a random subset of quarters destroys the serial dependence the block bootstrap exists to capture and would understate the interval a short sample really produces — and re-running the same bootstrap on each gives the left panel of Figure 4. The half-width does fall with T, but more slowly than the parametric rate and at a rate that differs by model: regressing the log half-width on $\log T$ gives a slope of -0.365 pre-COVID and -0.129 on the COVID-excluded sample for the factor model against the AR benchmark, against a parametric -0.5; across models the slopes range from -0.486 for quantile gradient boosting to $+0.167$ for the historical mean, where the interval widens as the window lengthens. Over the same range the model confidence set holds steady at about nine of eleven arms (right panel). The observed range spans a factor of three and the windows overlap heavily, so we report the curve and make no extrapolation from it to the sample length that would be required.

5.2 How few quarters decide the comparison

The reason the interval is wide, and the reason it narrows reluctantly, is visible in the remaining columns of Table 9. The per-quarter loss difference against the AR benchmark has a kurtosis

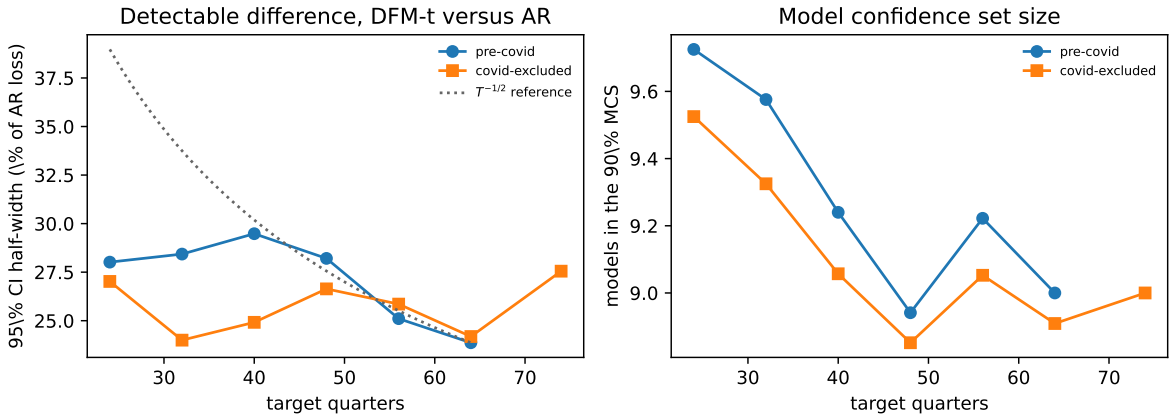

Figure 4: Interval half-width and model confidence set size against the number of target quarters, from contiguous windows of the pre-COVID and COVID-excluded samples at $mtq = 1$. The dotted line is the parametric rate anchored at the longest window, shown for comparison only.

between 11 and 20 for every arm that competes with it, and the three largest quarters supply 71 to 80 per cent of its sum of squares. For the factor model those three quarters are 2009:Q1, 2008:Q4 and 2009:Q4; for the structured estimator, 2009:Q2, 2009:Q1 and 2009:Q4; for quantile gradient boosting, 2009:Q3, 2009:Q4 and 2012:Q4.

The arms that lose heavily show the same concentration for the opposite reason. For the capacity-matched raw-lag arm the three decisive quarters are 2018:Q4, 2018:Q2 and 2019:Q2, and for the ADL bridge they fall in 2018 and 2019 — episodes of the arm's own instability rather than of the business cycle. Either way, a handful of quarters carries the comparison.

5.3 What happens when those quarters are removed

Table 10 recomputes each gain with the influential quarters dropped. Removing the eight quarters of 2008 and 2009 — applied identically to every model, and not one of the pre-specified regimes — moves the point estimates substantially and in both directions.

Across the sixty model-cells, seven change sign. Their distribution is what matters. The structured estimator of this paper changes sign at all three pre-COVID origins, moving from -5.0 to $+10.0$ per cent at $mtq = 2$, from -3.7 to $+11.9$ at $mtq = 1$ and from -14.2 to $+14.3$ at $mtq = 0$. In the opposite direction, the factor model's advantage on the COVID-excluded sample reverses, from $+8.2$ to -1.6 per cent at $mtq = 2$ and from $+4.7$ to -8.7 at $mtq = 1$.

That the same operation dissolves this paper's negative result for its own estimator and the factor model's positive one is what makes the exercise informative rather than self-serving. But the size of the movement should be read against Section 5.1. Every one of the seven sign changes is smaller than the half-width of the interval in its own cell: the largest, 28.5 percentage points for the structured estimator at $mtq = 0$, sits inside a half-width of 40.3, and the 15.6-point change at $mtq = 1$ inside a half-width of 23.4. The movement is real arithmetic on the point estimates; it is not by itself evidence of a distinct crisis regime.

Table 10: Are these statements about the financial crisis? CRPS gain over the quarterly-updated AR at mtq = 1, pre-COVID, recomputed with influential target quarters removed.

Model	All 64	Drop top 1	Drop top 3	Drop 2008–09	
DFM-t	15.9	9.8	8.2	6.3	
Quantile gradient boosting	7.5	3.6	1.3	4.0	
Random walk	1.3	-8.1	-9.2	-18.4	$*$
Structured CRPS-MIDAS	-3.7	3.5	6.9	11.9	$*$
Historical mean	-7.5	-11.8	-6.8	-19.4	
QRNN-MIDAS (published)	-10.3	-15.5	-9.9	-27.0	
Combined bivariate ADL bridge	-18.4	-10.4	-0.3	-42.3	
Raw-lag (capacity matched)	-53.2	-39.9	-23.3	-85.2	
QRNN-MIDAS, all 126 series	-63.6	-56.3	-45.5	-95.6	
Raw-lag (unmatched)	-176.9	-155.6	-133.3	-250.3	

Gains are per cent of the AR benchmark's loss. $*$ marks a change of sign. Columns 3 and 4 drop the quarters with the largest absolute loss difference for that model; the last drops the eight quarters of 2008 and 2009 for every model alike. The structured estimator changes sign at all three origins, moving from below AR to roughly ten per cent above it, while the DFM's advantage in the COVID-excluded regime reverses. Neither the negative result for the neural arm nor the positive one for the factor model survives as a claim about ordinary quarters. Excluding the crisis is a post-hoc subsample, not one of the pre-registered regimes, and is reported only as a sensitivity.

5.4 Local predictive ability, and stability proper

Dropping 2008–2009 is a choice made by the analyst. The fluctuation test of Giacomini and Rossi (2010) puts the same question without that choice: it tracks local relative performance over a rolling window and asks whether it is constant at zero, so the data locate any instability. We take the critical values from the limiting distribution by simulation rather than from the published table, so that the number reported here is regenerated by our own code rather than transcribed; at $\mu = 1$, where the statistic reduces to $|N(0,1)|$, the simulation returns 1.976 against the exact 1.960. A size study of the whole pipeline on null data gives 0.034 at $\mu = 0.3$ and 0.046 at $\mu = 0.5$ against a nominal 0.05, so the test is if anything conservative.

The test finds almost nothing; Figure 5 plots the paths. At $\mu = 0.3$ it rejects in one of sixty model-cells and at $\mu = 0.5$ in six, against three expected by chance at the 5% level. Every rejecting cell is one in which the arm is worse than the AR benchmark in *every* window — a statement about level, not about stability. The factor model, the boosted quantile model and the structured estimator never reject at any window length, origin or regime, with $\max|F|$ between 1.25 and 2.78 against critical values of 2.81 and 3.08.

The null of this test, however, is zero local relative performance, so it mixes level with stability: a uniformly inferior arm rejects everywhere, and the rejections above are exactly of that kind. To target stability alone we rerun the statistic on the demeaned differential $d_t - \bar{d}$, whose null is constancy at the sample's own level. Demeaning changes the limiting law to a Brownian-bridge increment, so the critical values are re-simulated for that functional (2.64 and 2.12 at the two windows); and because these prove conservative at 64 quarters — empirical size below 0.01 on iid null data — we also test against finite-sample critical values calibrated on that null (2.35 and 1.91,

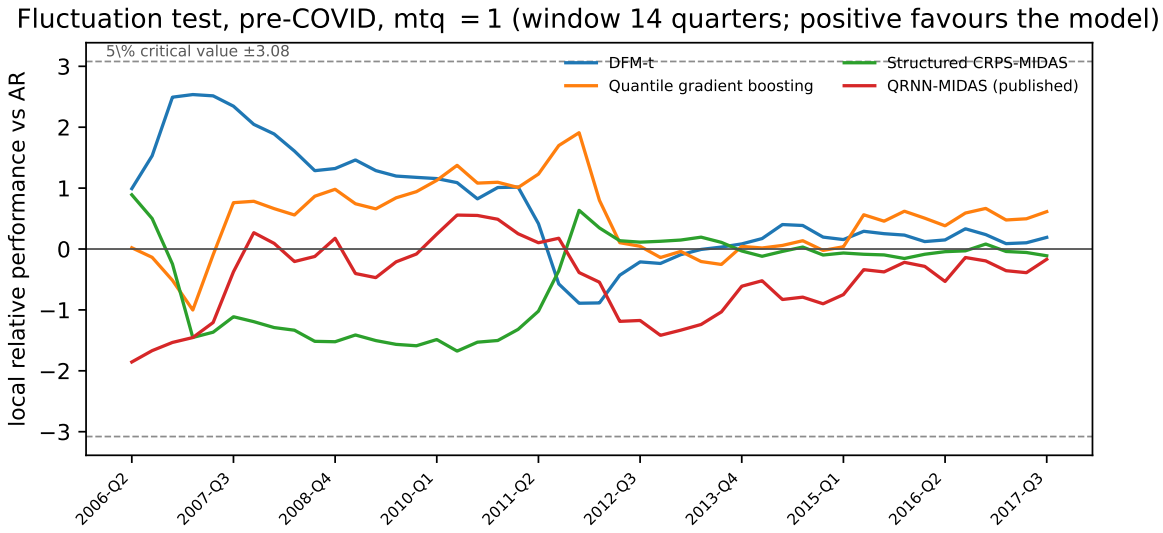

Figure 5: Local relative performance against the quarterly-updated AR benchmark over a rolling window of nineteen target quarters, pre-COVID, $mtq = 1$. Positive values favour the model. Dashed lines are the simulated 5% critical values.

size 0.05 by construction). Stability proper is rejected in *zero* of sixty model-cells under either set of critical values at either window. A non-rejection by one such test is weak evidence on its own; the same zero across every cell, window and calibration is at least consistent, and it agrees with the sign pattern of Section 5.3.

The two results of this section are therefore consistent rather than contradictory. Removing influential quarters moves the point estimates enough to reverse signs, and the proper test cannot detect instability in relative predictive ability — because the swings, large as they look, are smaller than the resolution established in Section 5.1. This sample determines neither the ranking among the retained arms nor the stability of that ranking — the arms it can exclude, it excludes at every scheme we tried. That is the limit within which every claim in this paper, including its own estimator's failure, must be read.

6 Conclusion

A published density neural MIDAS model, implemented to its authors' specification, does not reach a quarterly-updating autoregression at any within-quarter origin in this sample. Neither does the structured estimator this paper set out to evaluate. That much is stable: the sign holds in all eighteen pre-specified headline cells for each of four neural arms, across two regimes, three origins and three losses.

What the sample cannot do is support the converse. The model confidence set retains the historical mean and the random walk in every cell we examined, so no model here can be said to beat the benchmark either — including the two, a Student-t factor model and a quantile gradient

booster, whose point estimates are best. At 64 target quarters the design resolves differences of roughly twenty per cent of the benchmark's loss at the headline origin, where the largest point estimate is 15.9 per cent — 28.4 per cent across all eighteen cells, still inside its interval — so the null is a statement about resolution rather than about the models. Three of those 64 quarters carry most of the squared loss difference behind each competitive comparison; removing the crisis quarters changes seven signs in sixty, and a fluctuation test nonetheless finds no detectable instability, because the changes are smaller than the intervals they sit inside. The sample determines neither the ranking among the retained arms nor the stability of that ranking.

Between those two findings sits the argument of Section 4: each of the four design choices examined there moves what the comparison reports, and each is defensible. The published comparison we examine reports gains against QR-MIDAS and QRNN and contains no benchmark from outside those families. Adding one reverses the conclusion without contradicting the original arithmetic. Our own presentation is subject to the same reading, which is why the larger comparison set is reported as the headline even though the smaller one is more favourable to the models we propose.

Three limitations bound all of this. The evidence is one country, one target and 64 to 74 target quarters; the exclusions occur under a density score and, at the 90 per cent level, not under squared error; and the diverging classical benchmarks in Section 3.6 are facts about our implementations, so nothing here licenses a claim about the benchmarks used elsewhere.

Two questions are left open. The first is why a tree ensemble on fixed Almon summaries posts the second-best point estimate in the set while revising by almost exactly as much as the structured estimator, which posts a deficit. Three differences separate them and this sample cannot isolate any of the three. The second is what a comparison of this kind would need in order to decide anything. Lengthening the evaluation window within the range available here does not narrow the interval at the parametric rate, and the loss difference is carried by a few recessionary quarters, so more calendar time of the same kind is unlikely to be the answer. A second country would help less than it appears to, since the crises would largely coincide. We suspect the honest requirement is a comparison designed around the episodes that carry the information rather than around the quarters that happen to be available, and we do not have one.

Data availability statement

The data that support the findings of this study are openly available. These data were derived from the following resources available in the public domain: the Federal Reserve Bank of Philadelphia's Real-Time Data Set for Macroeconomists (Croushore and Stark, 2001), at `https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/real-time-data-set-for-macroeconomists`; and the FRED-MD monthly database (McCracken and Ng, 2016), at `https://research.stlouisfed.org/econ/mccracken/fred-databases/`. The code that constructs the evaluation panel and regenerates every exhibit and quoted number in the paper will be made available upon publication.

Table 11: First-available versus most-recent GDP. CRPS gain over the quarterly-updated AR benchmark, in per cent of AR loss, on the 83 target quarters where both vintages exist.

Model	mtq	First available	95% CI	Most recent	95% CI
DFM-t	2	12.8	[-8, 37]	10.6	[-9, 32]
DFM-t	1	15.9	[-5, 42]	12.6	[-8, 37]
DFM-t	0	2.7	[-19, 24]	5.6	[-13, 24]
Structured CRPS-MIDAS	2	-5.0	[-31, 14]	-21.2	[-53, 2]
Structured CRPS-MIDAS	1	-3.7	[-31, 15]	-20.2	[-55, 3]
Structured CRPS-MIDAS	0	-14.2	[-62, 17]	-26.6	[-78, 7]

Kim and Swanson (2018) report that this choice strongly affects model selection for real-time Korean GDP. Here the ranking is unchanged; the structured estimator's deficit against AR widens when it is scored against revised data.

A Data

A.1 Target

The target is quarter-on-quarter growth in US real GDP at an annual rate, $(Y_q/Y_{q-1})^4 - 1$, taken from the *first* vintage in which each quarter appears — the advance release, which in this panel falls exactly two months after the quarter closes, in all 861 scored origin-quarter rows — rather than from a later revised vintage. The backcast origin of Section 2.1 falls one month after the quarter closes and one month before that release, in every case. Quarterly vintages come from the Federal Reserve Bank of Philadelphia's real-time data set (Croushore and Stark, 2001).

Scoring the first release rather than the current one is a design choice with consequences. It matches what a forecaster is trying to predict at the time, and it avoids crediting a model for anticipating revisions that were not yet knowable; against that, it scores against a noisier target. Whether the choice matters here is reported in the paper rather than assumed. Table 11 rescores against the most recent vintage: the ordering of the two arms is unchanged at every origin, and the structured estimator's deficit against the benchmark widens, from 3.7 to 20.2 per cent at $mtq = 1$.

The evaluation covers target quarters from 2003:Q4 to 2026:Q2, of which the pre-COVID sample uses the 64 quarters of 2004–2019 and the COVID-excluded sample the 74 quarters of 2004–2025 excluding 2020–2022. The target has a standard deviation of 0.0182 on the pre-COVID sample. The two extreme quarters, 2020:Q2 and 2020:Q3, take values of -0.329 and $+0.331$, eighteen times the pre-COVID standard deviation; this is why the protocol declares full-sample rankings a contrast only and runs no test on the COVID quarters themselves. The sample ends at 2025:Q2 because the final estimation cutoff (2024) scores the six quarters 2024:Q1–2025:Q2, each earlier cutoff scoring the following two years; this is also why the COVID-excluded evaluation sample has $64 + 10 = 74$ quarters. The COVID exclusion applies to evaluation only: the expanding training windows of the later cutoffs retain those quarters once they are past.

A.2 Predictors

The headline monthly-vintage information set uses the FRED-MD monthly database (McCracken and Ng, 2016) at vintage-consistent releases: at each forecast origin, each of the 126 series enters at the vintage actually published by that date, with the publication lags of that date. Transformations follow the FRED-MD codes.

The robustness information set, full pseudo-real-time, keeps the origin timing but relaxes the vintage requirement, taking predictors at their current vintage. It is the looser of the two and is reported as a robustness check throughout. Neither design is real-time: vintages are historical and the computation is retrospective, so the correct description is pseudo-real-time.

The QRNN-MIDAS arm in its published specification uses six of the 126 series — industrial production, the consumer and producer price indices, M2, the oil price and the S&P 500 — as Xu et al. (2021) name for their US application. The auxiliary variant uses all 126.

A.3 Splits

Models are fitted at eleven expanding cutoffs at two-year intervals, 2004 through 2024. Within each cutoff the training window is split again, and all hyperparameter selection, early stopping and model choice use the inner validation fold; the outer block is scored once. Neural arms are fitted at five seeds and their predictive draws pooled.

B The revision contrast as an outcome-free statistic

For the factor model and the structured estimator on the same 64 pre-COVID target quarters, the CRPS difference at the three nowcast origins is -20.4, -23.2 and -17.4 per cent of the factor model's loss, with 95% intervals $[-74.3, 15.8]$, $[-86.3, 16.7]$ and $[-78.1, 21.3]$: wide, and all containing zero. The revision contrasts between the same two models, over the same quarters and under the same bootstrap, are 0.293 and 0.308 standard deviations of the target at the two within-quarter steps, with intervals $[0.206, 0.382]$ and $[0.193, 0.434]$: bounded away from zero. As a share of its own interval half-width the loss estimate is 0.45, 0.45 and 0.35 across the three origins, while the revision contrasts reach 3.3 and 2.6 — a sixfold to ninefold gap in precision on identical data. The same panel that cannot order the two models' accuracy orders their responsiveness without difficulty, and this appendix records the elementary reason.

Fix a target quarter and two adjacent origins, and write $y = \mu + \varepsilon$, where μ is the predictable component of the target given the later origin's information and ε the remainder, with variance σ_ε^2, independent across quarters and of the forecasts. For models $m \in \{A, B\}$ let f_m denote the later-origin forecast and r_m the revision between the two origins.

Proposition. (i) Any statistic built from the revision paths $\{r_{A,t}, r_{B,t}\}_{t \le T}$ alone — in particular the revision-size contrast of Section 3.3, whose scaling by the target's full-sample standard deviation is a constant held fixed under the bootstrap — has a sampling distribution that does not involve ε:

its variance is governed by the fourth moments of the revisions and is free of σ_ε^2 and of the tails of ε. (ii) Under squared error, writing $g_t = f_{A,t} - f_{B,t}$ and $d_t = \mu_t - (f_{A,t} + f_{B,t})/2$, the loss differential is $\Delta L_t = 2\, g_t(\varepsilon_t + d_t)$, so the mean loss difference satisfies

$$\mathrm{Var}\big(\overline{\Delta L}\big) \;=\; \frac{4}{T}\Big(\sigma_\varepsilon^2\, \mathrm{E}\big[g_t^2\big] + \mathrm{Var}(g_t d_t)\Big) \;\geq\; \frac{4\,\sigma_\varepsilon^2\, \mathrm{E}\big[g_t^2\big]}{T},$$

and its finite-sample interval inherits the tails of $g_t\varepsilon_t$, which grow with the kurtosis of ε and with the concentration of g_t on few dates.

Proof. Expand $(y - f_B)^2 - (y - f_A)^2 = (f_A - f_B)(2y - f_A - f_B) = 2g(\varepsilon + d)$; take the variance of the mean under independence of ε from the forecasts and across quarters, and drop the second, nonnegative term. Part (i) is immediate, since revisions are functions of the forecast paths only. □

Two boundaries of the claim. The slope β of Section 2.5 regresses a realised error on the revision and therefore does use the outcome; only the size contrast is outcome-free. And the display takes the pairs (g_t, d_t) independent across quarters, matching the setup above; under serial dependence the second term becomes a long-run variance and the lower bound is unchanged. The proof is for squared error; we use it as the transparent case for a pattern the CRPS numbers above exhibit, not as a theorem about CRPS.

Part (ii) says that an accuracy comparison must pay for the outcome: however smooth the two forecast paths, the loss differential carries $\sigma_\varepsilon^2\, \mathrm{E}[g^2]$, and when a few quarters carry most of $g\varepsilon$ — the kurtosis of the per-quarter loss difference is between 11 and 20 in this panel (Table 9) — the effective information shrinks further. Part (i) says the responsiveness comparison pays nothing for the outcome, because it never touches it. The two facts together are the precision gap in the numbers above. Part (i) is also the diagnostic's stated limit (Section 2.5): a statistic free of the outcome is necessarily silent about accuracy, which is why the paper uses it only for the responsiveness question.

The ingredients are old. Nordhaus (1987) tests a single forecaster's rationality through the autocorrelation of its revisions, Isiklar and Lahiri (2007) read the information content of survey horizons from revision behaviour, and Patton and Timmermann (2012) derive multi-horizon bounds for one forecast path. Using the revision contrast across models, with the outcome-independence above as the reason it separates where loss tests cannot, appears to be undocumented, though the algebra is elementary; we claim convenience for this design, not depth.

C Model specifications, densities and scoring conventions

This appendix records how each arm's predictive distribution is built and scored, and two conventions a replicator must know.

Table 12: Structured estimator: CRPS of each training seed's predictive distribution against the pooled five-seed mixture the paper scores. Pre-COVID, target-quarter-equal.

Origin	Seed range	Seed mean	Seed sd	Pooled mixture
$mtq = 2$	[0.0095, 0.0187]	0.0143	0.0033	0.0110
$mtq = 1$	[0.0093, 0.0187]	0.0142	0.0034	0.0109
$mtq = 0$	[0.0095, 0.0189]	0.0143	0.0034	0.0110

The pooled mixture reproduces the master panel's number exactly and, by the convexity of CRPS in the predictive distribution, cannot score worse than the seed average; here it is about 23 per cent better. Evaluating seeds individually would therefore make the arm look worse, not better, so the mixture convention is the generous one and is disclosed as the estimand.

C.1 Predictive distributions and the CRPS estimator

Every benchmark arm archives 2,000 predictive draws per forecast. The Student-t parametric arms — the AR benchmark, the factor model and the Student-t MIDAS specification — additionally admit a closed-form Student-t CRPS, and for those three the sealed headline score is the analytic value, with the draw-based estimator archived alongside; for every other arm the sealed score is a draw-based U-statistic estimator of the CRPS (Gneiting and Raftery, 2007), the exact estimator being recorded per row in the archive. Where both are available the analytic and draw-based values differ by sampling noise below 10^{-3} at this draw count. The QRNN-MIDAS adapter archived scores and PITs but not draws, which is why the quantile-score table below omits those two arms.

The structured estimator maps the monthly panel through a learned aggregation to a factor summary of dimension $r = 4$ (with $r_d = 8$ draw-noise channels), then through a neural head with a spectral-norm constraint and a bounded output transform (tanh scaled to the target's range). It is trained by CRPS on simulated draws (learning rate 0.002, at most 800 steps, 20 draws per step), with inner-validation CRPS evaluated every 10 steps and early stopping at patience 5; each of five seeds then generates 400 predictive draws, and the paper scores the pooled 2,000-draw mixture.

C.2 Seed pooling and its estimand

Pooling the seeds evaluates a mixture distribution in which initialisation uncertainty is part of predictive uncertainty. Table 12 reports the seed-level decomposition: seeds differ by a factor of two in CRPS, and the pooled mixture is about 23 per cent better than the seed average, as the convexity of CRPS guarantees it must weakly be. The convention is therefore generous to the neural arm — scored seed by seed it would trail the benchmark by roughly 35 per cent at $mtq = 1$ rather than the headline 3.7 — and the paper's negative conclusion survives either convention.

One arm follows a different convention, which we disclose rather than repair: the sealed panel scores the unmatched raw-lag arm by averaging per-seed losses instead of pooling draws. Its sealed relative CRPS of 2.769 therefore embeds a pooling penalty; on the mixture convention of every other neural arm its quarter-equal CRPS at $mtq = 1$ would be 0.0157 rather than 0.0291, still well above the benchmark. No conclusion changes under either convention, and the sealed numbers are retained throughout.

Table 13: Pinball loss relative to the AR benchmark at the five quantile levels of the density MIDAS literature: pre-COVID, $mtq = 1$, target-quarter-equal.

Model	$\tau = .1$	$\tau = .3$	$\tau = .5$	$\tau = .7$	$\tau = .9$
DFM-t	0.638	0.891	0.953	0.832	0.781
Quantile gradient boosting	0.921	0.992	1.094	0.867	0.591
Random walk	0.687	0.922	1.117	1.089	0.963
Historical mean	0.981	1.157	1.239	1.097	0.781
Structured CRPS-MIDAS	1.049	1.017	1.193	0.983	0.736
Combined bivariate ADL bridge	1.027	1.135	1.301	1.225	1.153
Raw-lag (capacity matched)	1.731	1.390	1.393	1.443	1.543
Raw-lag (unmatched)	1.583	1.672	1.381	1.324	1.193

Values below one favour the model; quantiles are read from each arm's archived predictive draws. The two QRNN-MIDAS arms are absent because their adapter archived scores and PITs but not draws, the only arms for which that is so.

C.3 Quantile scores and calibration

Table 13 scores the arms under the pinball loss at the five quantile levels used by the density MIDAS literature, read from the archived draws. Table 14 reports calibration. The two tables answer the question a scalar score cannot: *which aspect* of a predictive distribution fails. The QRNN-MIDAS arms' central 80 per cent intervals cover about half their nominal rate — densities far too narrow — while the structured estimator is close to nominal coverage, its failure being location and inertia rather than dispersion, consistent with the drift decomposition of Section 3.3. The AR benchmark over-covers.

C.4 Sensitivity and accounting

Table 15 varies the resampling scheme — block lengths 3, 4, 5 and 8, a stationary bootstrap with mean block five, and a segment-aware variant that never draws a block across the COVID join of the COVID-excluded sample. The headline interval barely moves and the excluded arms of the 90 per cent confidence set are identical under every scheme. Table 17 lists every row removed by the pre-specified divergence filter, by arm and origin, including a single flagged factor-model row at a secondary origin. Table 18 reports the third pre-specified loss, absolute error, completing the eighteen headline cells in exhibit form.

Table 16 asks the complementary question to Table 15: not whether the interval moves with the resampling scheme, but whether the percentile interval covers at its nominal level in this design at all. From 2,000 Monte Carlo replications of the paper's bootstrap at $T = 64$: under an independent Gaussian differential coverage is 0.927; Student-t tails at 5, 3 and 2.5 degrees of freedom leave it between 0.912 and 0.916; first-order autocorrelation of 0.3 costs a further two to three points; and resampling the observed per-quarter loss pairs — so that the concentration of Section 5.2 reappears in every draw — gives 0.910 to 0.923 for the factor-model comparison and 0.872 to 0.875 for the structured arm. The empirical-shape draws are independent across quarters, so dependence

Table 14: Calibration of the predictive distributions: empirical coverage of the central 50% and 80% intervals and the mean PIT. Pre-COVID, $mtq = 1$, 64 target quarters per arm.

Model	Coverage 50%	Coverage 80%	Mean PIT
AR-t (quarterly update)	0.750	0.938	0.482
DFM-t	0.609	0.828	0.478
Quantile gradient boosting	0.344	0.734	0.452
Random walk	0.500	0.875	0.539
Historical mean	0.516	0.797	0.603
Structured CRPS-MIDAS	0.500	0.781	0.434
QRNN-MIDAS (published)	0.297	0.500	0.378
QRNN-MIDAS, all 126 series	0.250	0.500	0.549
Combined bivariate ADL bridge	0.406	0.734	0.505
Raw-lag (capacity matched)	0.250	0.422	0.509
Raw-lag (unmatched)	0.375	0.656	0.435

Nominal values are 0.50, 0.80 and 0.50. PITs come from archived draws, and from the sealed panel column for the two QRNN-MIDAS arms. The AR benchmark over-covers (its densities are too wide); the QRNN-MIDAS arms cover half their nominal 80% interval (densities far too narrow); the structured estimator is close to nominal, so its failure is location and inertia, not dispersion.

Table 15: Sensitivity of the headline interval and the 90% model confidence set to the resampling scheme, at $mtq = 1$.

Scheme	DFM-t vs AR, pre-COVID	DFM-t vs AR, COVID-excl.	MCS pre	MCS excl.
circular blocks, length 3	[-4.1, 43.1]	[-21.8, 32.6]	9	9
circular blocks, length 4	[-4.5, 42.8]	[-21.9, 33.3]	9	9
circular blocks, length 5 (headline)	[-5.1, 41.7]	[-22.0, 32.3]	9	9
circular blocks, length 8	[-4.9, 41.9]	[-21.4, 32.5]	9	9
stationary bootstrap, mean 5	[-3.1, 40.4]	[-20.8, 31.0]	9	9
segment-aware, length 5	–	[-17.1, 30.3]	–	9

Intervals are 95% percentile intervals for the CRPS gain in per cent of the AR benchmark's loss; MCS columns count retained arms of eleven. The segment-aware variant never draws a block across the 2019–2023 join of the COVID-excluded sample and therefore exists only there. The excluded arms are identical across every scheme.

sensitivity remains the preceding table's job. The under-coverage is modest, is driven by the small sample and the blocking rather than by the tails, and runs in the only safe direction here: wider, correctly calibrated intervals would contain zero wherever the reported ones do.

D Why a diverging arm freezes the model confidence set

Section 3.6 reports that with the diverging classical arms left in, the 90 per cent model confidence set eliminates nothing. The panel shows why in an unusually clean form. In the first round of the procedure the studentized statistics of all thirteen arms collapse to two values: $+1.46$ for the diverging Student-t MIDAS arm and -1.46 for every other arm at $mtq = 2$, and ± 1.37 and ± 1.38 at the other two origins. The bootstrap p-values attached to these maxima are 0.17, 0.20 and 0.16 — far above the 0.10 threshold — so no elimination can begin. Nor is the collapse an artefact of

Table 16: Monte Carlo coverage of the 95% percentile interval under the paper's bootstrap (circular blocks of length five, $T = 64$): 2,000 replications of 1,999 bootstrap draws each.

DGP	Dependence / origin	True value	Coverage (MC s.e.)	Mean width
Panel A: mean of $T = 64$ simulated loss differentials, true mean zero				
Normal	independent	0	0.927 (0.006)	0.46
Normal	AR(1), $\rho = 0.3$	0	0.895 (0.007)	0.61
Student-t, 5 df	independent	0	0.915 (0.006)	0.59
Student-t, 5 df	AR(1), $\rho = 0.3$	0	0.881 (0.007)	0.78
Student-t, 3 df	independent	0	0.912 (0.006)	0.76
Student-t, 3 df	AR(1), $\rho = 0.3$	0	0.888 (0.007)	1.00
Student-t, 2.5 df	independent	0	0.916 (0.006)	0.91
Student-t, 2.5 df	AR(1), $\rho = 0.3$	0	0.894 (0.007)	1.17
Panel B: relative CRPS gain, quarters resampled from the observed pre-COVID panel				
DFM-t	$mtq = 2$	12.8%	0.923 (0.006)	27.6pp
Structured CRPS-MIDAS	$mtq = 2$	-5.0%	0.875 (0.007)	33.7pp
DFM-t	$mtq = 1$	15.9%	0.910 (0.006)	29.0pp
Structured CRPS-MIDAS	$mtq = 1$	-3.7%	0.872 (0.007)	33.4pp
DFM-t	$mtq = 0$	2.7%	0.916 (0.006)	33.8pp
Structured CRPS-MIDAS	$mtq = 0$	-14.2%	0.872 (0.007)	45.8pp

Panel A simulates mean-zero differentials and records how often the interval covers zero; widths are in innovation units. Panel B draws 64 quarters iid with replacement from the observed per-quarter (model, AR) CRPS pairs, so the loss concentration of the real panel reappears in every draw; the true value is the full-population gain and widths are in percentage points of the AR benchmark's loss. Nominal coverage is 0.95 throughout.

Table 17: Rows flagged by the pre-specified divergence filter, by model and origin (monthly-vintage information set, all scored rows).

Model	$mtq = -1$	$mtq = 0$	$mtq = 1$	$mtq = 2$	$mtq = 3$	$mtq = 4$	$mtq = 5$	$mtq = 6$	$mtq = 7$	$mtq = 8$
DFM-t	0	0	0	0	1	0	0	0	0	0
MIDAS-t	3	10	8	11	8	6	9	3	2	2
Quantile MIDAS	11	15	12	14	17	17	16	8	8	12

The filter flags any row whose predicted mean exceeds a threshold the target's scale cannot reach. Every arm not listed has zero flagged rows; the single DFM-t row (a secondary origin) is disclosed here and included in every filtered variant exactly as the rule dictates.

Table 18: Absolute error relative to the quarterly-updated AR benchmark, by nowcast origin: pre-COVID, monthly-vintage information set, target-quarter-equal.

Model	$mtq = 2$	$mtq = 1$	$mtq = 0$
DFM-t	1.002	0.956	1.058
Quantile gradient boosting	1.056	1.026	1.097
Random walk	1.305	1.119	1.183
Historical mean	1.195	1.189	1.269
Structured CRPS-MIDAS	1.150	1.127	1.204
QRNN-MIDAS (published)	1.169	1.146	1.278
Combined bivariate ADL bridge	1.562	1.300	2.370
Raw-lag (capacity matched)	1.687	1.642	1.772
Raw-lag (unmatched)	2.431	2.324	2.459

Values below one favour the model. This is the third of the three pre-specified losses entering the eighteen headline cells; CRPS and squared error appear in the main text.

the particular magnitudes: in a synthetic panel with one arm inflated on three dates, raising the inflation from ten times the loss scale to 10^8 times moves that arm's statistic only from 1.75 to 1.94, where it stays.

The mechanism is the classical behaviour of self-normalized sums (Logan et al., 1973). Suppose one arm's losses are of size M on k dates and negligible elsewhere, with M large relative to every other arm. Its average excess over the set mean then grows linearly in M; but so does the bootstrap standard error of that excess, because the resampling variation is carried by the very dates that carry the mean. Their ratio — the studentized statistic — converges, as M grows, to a limit that depends only on the outlier configuration, the block structure and the number of arms, and not on M at all. The shared centring propagates the same freeze to every other arm's statistic, and the bootstrapped null distribution, being self-normalized in the same way, freezes with them. Whether anything is eliminated is then decided entirely by where these limits happen to fall, and in this panel they fall inside the acceptance region at the 90 per cent level and outside it at the 75 per cent level, which is why the diverging arms eliminate themselves there and only there (Section 3.6).

The practical consequence is the one stated in Section 3.6: a pre-specified screen for diverging arms is not cosmetic but a condition for the procedure to discriminate at conventional levels. The limit argument above holds the sample and the outlier configuration fixed and lets the divergence grow, so within that frame no increase in bootstrap depth or divergence magnitude changes anything; how the freeze behaves along joint asymptotics in the sample size, the contamination frequency and the block structure is a question we leave open. The underlying principle is old — self-normalized statistics do not grow with the contamination, which is also what makes t-statistic approaches robust in the opposite, intended direction (Ibragimov and Müller, 2010) — but its consequence for the model confidence set appears not to have been recorded. As with Appendix B, we claim convenience for this design rather than depth, and we present the argument as a mechanism sketch supported by the exact collapse above rather than as a theorem.

References

Aastveit, K. A., Gerdrup, K. R., Jore, A. S., and Thorsrud, L. A. (2014). Nowcasting GDP in real time: A density combination approach. *Journal of Business & Economic Statistics*, 32(1):48–68.

Adrian, T., Boyarchenko, N., and Giannone, D. (2019). Vulnerable growth. *American Economic Review*, 109(4):1263–1289.

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics*, 25(2):177–190.

Andreou, E., Ghysels, E., and Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, 158(2):246–261.

Babii, A., Ghysels, E., and Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, 40(3):1094–1106.

Bánbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). Now-casting and the real-time data flow. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2A, pages 195–237. Elsevier, Amsterdam.

Clements, M. P. and Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States. *Journal of Business & Economic Statistics*, 26(4):546–554.

Croushore, D. and Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics*, 105:111–130.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Foroni, C., Marcellino, M., and Schumacher, C. (2015). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A*, 178(1):57–82.

Ghysels, E., Sinko, A., and Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26(1):53–90.

Giacomini, R. and Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, 25(4):595–620.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., and Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5):920–964.

Grant, A., Mrazik, M., and Satchell, S. (2026). Evaluating forecasts at multiple horizons: An extension of the Diebold–Mariano approach. *Journal of Forecasting*. Early view.

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Ibragimov, R. and Müller, U. K. (2010). t-statistic based correlation and heterogeneity robust inference. *Journal of Business & Economic Statistics*, 28:453–468.

Isiklar, G. and Lahiri, K. (2007). How far ahead can we forecast? Evidence from cross-country surveys. *International Journal of Forecasting*, 23(2):167–187.

Kim, H. H. and Swanson, N. R. (2018). Methods for backcasting, nowcasting and forecasting using factor-MIDAS: With an application to Korean GDP. *Journal of Forecasting*, 37(3):281–302.

Kuzin, V., Marcellino, M., and Schumacher, C. (2011). MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting*, 27(2):529–542.

Logan, B. F., Mallows, C. L., Rice, S. O., and Shepp, L. A. (1973). Limit distributions of self-normalized sums. *Annals of Probability*, 1(5):788–809.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, 34(4):574–589.

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., and Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, 39(1):98–119.

Németh, K. (2026). GDP nowcasting with artificial neural networks: How much does long-term memory matter? *Journal of Forecasting*. Early view.

Nordhaus, W. D. (1987). Forecasting efficiency: Concepts and applications. *Review of Economics and Statistics*, 69(4):667–674.

Patton, A. J. and Timmermann, A. (2012). Forecast rationality tests based on multi-horizon bounds. *Journal of Business & Economic Statistics*, 30(1):1–17.

Politis, D. N. and Romano, J. P. (1992). A circular block-resampling procedure for stationary data. In LePage, R. and Billard, L., editors, *Exploring the Limits of Bootstrap*, pages 263–270. Wiley, New York.

Richardson, A., van Florenstein Mulder, T., and Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, 37(2):941–948.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

Xu, Q., Liu, S., Jiang, C., and Zhuo, X. (2021). QRNN-MIDAS: A novel quantile regression neural network for mixed sampling frequency data. *Neurocomputing*, 457:84–105.