Research Square

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

Neural Forecasting of Daily Commodity Futures Returns Against a Zero-Return Benchmark

Navid Goodarzi

navidgoodarzi@ut.ac.ir

University of Tehran

Research Article

Keywords: forecast evaluation, neural forecasting, commodity futures, return predictability, global training, economic value

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11447734/v1

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

Additional Declarations: No competing interests reported.

Neural Forecasting of Daily Commodity Futures Returns Against a Zero-Return Benchmark

Navid Goodarzi

University of Tehran, Tehran, Iran

Corresponding author: navidgoodarzi@ut.ac.ir

Abstract

Machine-learning forecasts of financial returns are usually judged by whether they beat other models, but in low signal-to-noise settings such comparisons can mislead. We propose an evaluation protocol for neural return forecasts that combines four elements: a zero-return benchmark with loss-based block-bootstrap tests and confidence bounds on the panel mean out-of-sample R^2; an exact decomposition of R^2 into squared forecast–return alignment and an amplitude penalty; calibration that uses only past data; and the certainty-equivalent return of a mean–variance investor after transaction costs. We apply it to N-BEATS, N-HiTS and N-BEATSx on daily returns of 19 commodity futures from 2015 to 2026. No neural model improves significantly on zero, and the upper 95% bound on the mean daily R^2 is 0.16%. Training one network on all commodities raises the mean R^2 by 3.32 percentage points relative to separate training, yet almost all of this gain comes from smaller forecasts rather than a detectable improvement in alignment, and past-only calibration reduces it to 0.06 points. Sector-based pooling does no better than random grouping. Several forecasts raise the investor's certainty-equivalent return over a historical-mean benchmark before costs, but none beats equal-weight buy-and-hold, the gains vanish at one-way costs above about 4 basis points, and the training strategy with the lowest loss is not the one with the highest certainty-equivalent return. Lower forecast loss relative to another model should therefore not be read as evidence of predictability or of economic value.

Keywords: forecast evaluation; neural forecasting; commodity futures; return predictability; global training; economic value

1. Introduction

Daily commodity futures returns are difficult to forecast, despite the growing range of nonlinear methods available to researchers. Futures prices reflect much of the information available to market participants, while risk premia are small relative to day-to-day price movements. That makes a zero-return forecast a demanding benchmark, though market efficiency does not require the conditional mean to be exactly zero. Most documented commodity signals work at longer horizons [1, 2]. Angelidis, Sakkas and Tessaromatis [3], for example, find that cross-sectional models based on basis and momentum outperform time-series specifications for monthly returns, with pooled out-of-sample R^2 of about 1% relative to the historical mean. It is less clear whether neural networks can improve forecasts of daily returns. Beating another fitted model is not enough: a neural network can win that comparison and still lose to predicting zero.

Commodity markets also differ in ways that could matter for joint training. Storage costs, inventories and convenience yields affect futures pricing [4], while hedging pressure and the positions of market participants bear on risk premia [5, 6]. Energy contracts face supply and geopolitical shocks; metals respond to industrial demand and dollar movements; crops are exposed to weather and harvest cycles. These differences make sector pooling plausible, but not necessarily useful. Training on related markets increases the number of observations while potentially combining distinct return processes. Even closely related contracts, such as soybeans, soybean oil and soybean meal, differ in forecastability in Angelidis et al. [3]. Global models can forecast well even across dissimilar series [7], so whether pooling helps here is an empirical question.

A second concern is the size of the forecasts. Daily return variation is large relative to plausible changes in expected returns, so a model can incur substantial squared-error loss by predicting too

much movement. Shrinkage toward zero has a long history in return forecasting [8, 9]. In the commodity setting, Angelidis et al. [3] report Mincer–Zarnowitz slopes well below one for their time-series models and less variable forecasts from their more accurate cross-sectional models. This raises a specific question for neural forecasting: does a lower loss reflect better alignment with realized returns, or simply forecasts of smaller magnitude? The answer matters when comparing training strategies. Even an improvement in forecast loss would not, by itself, establish an investable futures strategy once transaction and roll costs are considered.

We examine these issues using daily returns on 19 commodity futures in energy, metals, and agriculture, evaluated out of sample from January 2015 to June 2026. This differs from Angelidis et al. [3] in the horizon (daily rather than monthly), the information used (past returns and, for one model, macroeconomic variables, rather than basis and momentum characteristics), the models (neural networks rather than regularized linear regressions) and the benchmark (a zero-return forecast rather than the historical mean). The neural models are N-BEATS and N-HiTS, and N-BEATSx with macroeconomic predictors. The conventional benchmarks are an AR(1), the historical mean, and gradient-boosted trees. We compare neural models trained separately for each commodity with models trained across all commodities and across groups of commodities, and we pay particular attention to the zero-return forecast. We use block-bootstrap loss comparisons for non-nested forecasts and Clark–West adjustments in the applicable nested linear comparisons. We also investigate the finite-sample behavior of forecast tests under heavy-tailed returns, since nominal significance levels need not provide reliable inference in this setting.

The results provide no statistically significant evidence that the neural models improve on the zero-return forecast at the daily horizon. The raw neural forecasts all have negative mean out-of-sample R^2 against the zero forecast, and the 95% upper confidence bound for the panel-mean R^2 is below 0.16% for every model examined. There are nevertheless substantial differences among the neural training strategies. Models trained across all commodities have much lower loss than models trained separately for each commodity, and the difference grows when the separate models use the hyperparameters of the pooled model. Yet neither approach beats the zero-return forecast.

To investigate this result, we separate out-of-sample R^2 into a squared forecast–return alignment component and a penalty associated with forecast amplitude. Almost all of the raw loss difference between pooled and separate training appears in the amplitude penalty, and the alignment component does not differ significantly between them. When we rescale forecasts using only information available before each prediction, the gap in mean R^2 between the two approaches falls from 3.32 to 0.06 percentage points and is no longer statistically significant. This result requires caution because the rescaling procedure sets more than half of the forecasts to zero. Grouping commodities by economic sector, the grouping that commodity economics would suggest, did not improve forecasting performance relative to random groupings. These comparisons were designed after the main study and are exploratory.

The size of the bound deserves a comment. A daily R^2 of 0.16% is not necessarily negligible in economic terms. As an illustration, a single asset whose predictable component has a daily R^2 of 0.16% would have a Sharpe ratio of $\sqrt{R^2/(1-R^2)}$ per day, or about 0.63 when annualized with $\sqrt{252}$, before costs (see [8] for the mean–variance framework). The bound applies to the panel mean of the commodity-level R^2 values and not to any single commodity, so this figure shows the scale of 0.16% and is not a Sharpe ratio that the models could achieve. The bound therefore limits the predictability that the data leave room for, but it does not rule out predictability that would matter to an investor. Our conclusion is that the neural models do not deliver it, not that none exists.

The paper makes two contributions. The first is an evaluation protocol for machine-learning forecasts of financial returns in low signal-to-noise settings, which any model can be passed through. It has four parts: a comparison with a zero-return forecast using loss-based block-bootstrap tests and confidence bounds on the panel mean out-of-sample R^2, with a simulation check of test size under heavy-tailed returns; an exact decomposition of R^2 into squared forecast–return alignment and an amplitude penalty; calibration of the forecasts that uses only past data; and the certainty-equivalent return of a mean–variance investor after transaction costs. None of these parts is new on its own. Their combination answers questions that a

ranking of models by loss cannot: whether a model beats the natural benchmark at all, whether a relative gain reflects more information or smaller forecasts, and whether either is worth anything to an investor. The second contribution is the application of this protocol to neural forecasts of daily returns on 19 commodity futures, under local, global and sector-level training. No model beats the zero forecast, and the bound on the mean daily R^2 is 0.16%. The large loss differences between training strategies are concentrated in forecast amplitude rather than in a detectable difference in alignment, which extends to neural networks and daily data the evidence on forecast scale that exists for linear models at the monthly horizon. In economic terms the ranking of training strategies by loss is not the ranking by certainty-equivalent return, no strategy beats equal-weight buy-and-hold, and the gains over a historical-mean investor vanish at one-way costs of about 1.5 to 4.3 basis points. These are descriptive results: we do not identify why pooled training changes forecast scale.

Section 2 reviews the literature, Section 3 sets out the data and tests, and Section 4 presents the results. Sections 5–7 discuss the findings, limitations and conclusions.

2. Literature Review

2.1. Predictability in commodity futures returns

Evidence of return predictability in commodity futures depends heavily on the target, forecast horizon and information set. Studies using inventories, risk-premium proxies or macroeconomic variables sometimes outperform simple benchmarks; others find little out-of-sample improvement. Those results cannot be transferred directly to a daily forecasting exercise with a different benchmark.

Cotter, Eyiah-Donkor and Potì [10] combine forecasts from 28 potential commodity, financial and macroeconomic predictors and find out-of-sample predictability of commodity futures excess returns. Predictability is countercyclical, the combination forecasts predict future economic activity, and the authors interpret them as state variables in the intertemporal asset pricing framework of Merton [11], which links the predictability to time-varying risk premia.

The evidence for individual commodities is more complicated. Angelidis, Sakkas and Tessaromatis [3] examine 37 futures markets, comparing conventional time-series models with forecasts based on cross-sectional commodity characteristics. Basis, momentum, and basis-momentum prove more useful than many of the financial and macroeconomic variables used in their time-series specifications. Their cross-sectional models, including regularized regressions, produce positive out-of-sample R^2 values for many contracts and improve the economic performance of forecast-based portfolios.

Cotter et al. combine forecasts from a broad set of predictors; Angelidis et al. instead exploit observed differences in commodity characteristics. Both use information that the networks studied here do not see, such as term-structure, cross-sectional and macroeconomic variables, and neither establishes how neural networks fitted to daily return histories would fare against predicting zero rather than the historical mean.

2.2. Machine learning and the use of information across markets

Wang and Zhang [12] study commodity futures return prediction with machine learning models and commodity-specific and macroeconomic predictors for 22 commodities. Their models have lower out-of-sample prediction errors than AR(1) and forecast-combination benchmarks for most of the contracts, and a monthly long–short strategy built on LightGBM forecasts beats the linear benchmarks on annual return, Sharpe ratio and maximum drawdown. Performance still differs across contracts. A method that works well for one commodity need not perform equally well for another.

For neural forecasting, the practical choice is between fitting one network per commodity and sharing parameters across several markets. Separate models can capture contract-specific patterns but have less training data. Joint estimation offers a larger sample, which may help with weak signals [7], at the cost of combining observations generated by different economic shocks. The inventory news relevant to crude oil, for instance, need not resemble the information driving wheat prices.

The work of Angelidis et al. is relevant here, although its cross-sectional models should not be

confused with pooled neural training. Their forecasts use differences in observed commodity characteristics to estimate expected returns. A global neural model instead learns shared parameters from several return series. Successful cross-sectional prediction does not guarantee that joint estimation will help, particularly if the commodities being pooled share little useful predictive structure.

Pooling can also change forecast scale without improving the information captured about future returns. If expected daily returns are close to zero, smaller forecasts may reduce squared-error loss even when their association with realized returns is no stronger. We therefore examine the forecasts as well as the resulting losses. The architectures are N-BEATS [13], N-HiTS [14] and N-BEATSx [15]; Gu, Kelly and Xiu [16] provide a broader application of neural networks to asset returns.

2.3. Forecast evaluation and calibration

Benchmark choice changes the question being tested. A model can beat AR(1) or a historical-mean forecast and still perform worse than zero. Out-of-sample comparisons have been central to the return-prediction literature precisely because apparent in-sample relationships often fail to survive evaluation on later data [17, 8].

The distinction is relevant when interpreting the commodity studies discussed above. Angelidis et al. [3] evaluate their results against historical-average forecasts, and Wang and Zhang [12] report comparisons with AR(1) and forecast combinations. Cotter et al. [10] work with combination forecasts. Those benchmarks are appropriate to the questions examined in each paper. None of these comparisons, by itself, establishes how a neural model trained on daily returns performs against a zero-return forecast.

Inference requires a comparable distinction. Diebold and Mariano [18] test differences in predictive accuracy; Clark and West [19] adjust for estimation noise in nested comparisons. The latter adjustment is not a general-purpose test for complex, non-nested forecasts. Heavy-tailed commodity returns further complicate conventional approximations, motivating our limited simulation check of test size.

Forecast magnitude introduces another difficulty. Under squared-error loss, a prediction that is directionally related to the realized return can still be inaccurate if its magnitude is too large. Angelidis et al. [3] find that the Mincer–Zarnowitz slopes of their time-series models are well below one, which they interpret as overfitting, and that their cross-sectional models produce less volatile forecasts than the time-series specifications. They list lower forecast volatility among the reasons for the cross-sectional advantage. Their findings provide a useful precedent: part of the difference between forecasting approaches can lie in the scale of their predictions.

The evidence reviewed here leaves open two questions for daily neural forecasts: whether sharing parameters across commodities improves accuracy against zero, and how much of any relative improvement comes from forecast scale. Section 3 describes our alignment–amplitude decomposition and the separate, past-only calibration exercise.

3. Methodology

3.1. Commodity futures data and return construction

The sample contains 19 Yahoo Finance continuous front-month futures series: four precious metals (gold, silver, platinum, palladium), copper, five energy products (WTI, Brent, natural gas, heating oil, gasoline), three grains (corn, wheat, soybeans), four softs (coffee, sugar, cotton, cocoa) and two livestock products (live cattle, lean hogs). The downloaded range is January 2000 to June 2026. Series start on different dates. The evaluation period starts in January 2015 and gives about 2,880 daily returns per commodity.

The price is the Yahoo Finance close with automatic adjustment switched off. We remove observations with non-positive prices and stale rows (zero volume and a close identical to the previous close). Returns are log changes of the remaining closes,

$$r_{i,t} = \ln P_{i,t} - \ln P_{i,t-1}, \qquad (1)$$

where $P_{i,t}$ is the close of commodity i on retained date t. When a stale row is dropped, adjacent retained prices are joined, so a return can span more than one calendar day. This applies, for example, to the non-positive WTI close of April

2020. The target at origin t is the cumulative log return over the next h retained days,

$$y_{i,t+h} = \sum_{s=1}^{h} r_{i,t+s}\,, \qquad (2)$$

with $h = 1$ as the primary horizon and $h = 5$ and $h = 21$ as secondary horizons.

Yahoo does not document how contracts are rolled in these series, and the series are not roll-adjusted. Price jumps at contract rolls can therefore enter the measured returns. We call them commodity futures price-series returns, and we do not treat them as returns on a rolled contract-level position. A data audit flags returns more than ten robust standard deviations from the median as possible roll or data errors; flagged days are kept in the main analysis. The WTI observation of April 2020 is the clearest case: removing it joins the two adjacent retained prices, so the return spans more than one day and may be artificial. As a check on the roll issue we repeat the main neural models on roll-clean energy returns built from EIA NYMEX contract prices, holding the first contract and moving to the second five trading days before the last trade day (Section 6). This check covers only the energy commodities and cannot validate the returns of all 19 series. Section 6 reports a sensitivity check that drops origins around volume jumps.

Information at the forecast origin is limited to prices and inputs dated on or before t. Return scaling, hyperparameter tuning and model estimation use only observations dated before the period being forecast, and the macro inputs are standardized with data from before 2015 (Section 3.2).

3.2. Forecasting models and training design

Benchmarks. The zero forecast is the main benchmark. A zero forecast is the conditional-mean forecast if returns have a zero conditional mean. It does not follow from market efficiency alone, since futures may carry a risk premium, and the historical-mean forecast is reported for that reason. The other benchmarks are the historical mean (the average return before the current month, times h), an AR(1) (a direct regression of the target on the current return with an intercept, re-estimated each month), gradient-boosted trees (XGBoost, with ten lagged returns as inputs and 40 tuning trials), and a naive forecast equal to h times the current return, which is kept only to illustrate how loss depends on forecast size.

Neural models. The neural models are N-BEATS and N-HiTS at all three horizons. At $h = 1$ we also estimate N-BEATSx with three historical exogenous inputs: daily changes in the U.S. dollar index, the VIX and the 10-year Treasury yield, aligned to commodity dates with a backward as-of merge. Each series is transformed to a log change (dollar index and VIX) or a first difference (Treasury yield), divided by its pre-2015 standard deviation, clipped at ±10, and set to zero when missing. The same-date alignment uses the latest macro close dated on or before the commodity date. A macro close can be recorded after the settlement of a given commodity contract on that date, so the exercise forecasts from end-of-day information and does not describe a rule that could be traded at the commodity close. Two post-hoc variants shift the macro inputs back, to levels dated strictly before the commodity date and to a one-row shift of the standardized changes. Each network uses squared-error loss and is trained for at most 1,500 steps with batches of 256 windows. Early stopping uses the last 300 observations of each training set as validation data, checks every 25 steps and stops after ten checks without improvement. N-BEATS and N-BEATSx have three generic (identity) stacks with one block each, and every block has two hidden layers of width w; N-HiTS has the same hidden layers with pooling kernel sizes of 2, 2 and 1, frequency downsampling factors of 4, 2 and 1, max pooling and linear interpolation, which are the NeuralForecast defaults. All blocks use ReLU activations and no dropout. These settings are the same for all commodities.

Use of AI tools. The large language model Claude (Anthropic) was used to help write and debug the Python code for the pipeline and to help draft and revise the text of the manuscript. Every result reported here comes from running that code. The author designed the study, checked the reported figures against the code outputs, reviewed and edited all text, and takes full responsibility for the content.

Local and global training. A *local* model is estimated separately for each commodity. A *global* model is estimated once on all 19 commodities. In both cases each return series is divided by the standard deviation of its own training sample, and forecasts are multiplied back to the original scale. The global models are run at $h = 1$.

Hyperparameters. Hyperparameters are chosen with 20 Optuna trials on pre-2015 data only. The search space is input window $\in \{5,10,20,40\}$, width $w \in \{8,16,32,64\}$, learning rate in $[10^{-4}, 3 \times 10^{-3}]$ (log scale) and weight decay in $[10^{-4}, 1]$ (log scale). The validation origins are the last 500 business days before 2015, the training data end before them, and the objective is the sum of squared errors relative to the zero forecast. Local models are tuned separately for each commodity. The global model is tuned once on the pooled data.

Re-estimation and ensembling. Parameters are re-estimated on the first calendar day of each month using only observations dated before that day, and they are held fixed for all forecast origins in the month. Neural forecasts are the average over three random seeds.

Leakage controls. Return scaling constants, early-stopping data and model parameters use only data dated before the month being forecast. The macro inputs are standardized with their standard deviations over the whole period before 2015, which includes the 500 business days used to validate hyperparameters, so for the tuning step this scale is not estimated strictly before the validation window; it does not use any data from the evaluation period. Targets enter a training set only if they are fully realized before the month starts. A no-look-ahead test perturbs all returns after a cutoff date and confirms that every forecast with an origin before the cutoff is unchanged, for the neural engine and all baselines.

3.3. Out-of-sample evaluation

Every trading day from January 2015 to June 2026 is a forecast origin. At origin t the model uses data through the close of t and the parameters estimated at the start of the current month. All models in a comparison are evaluated on the same origins; for comparisons across training strategies (Section 3.5) we use the origins common to all arms, which gives between 2,881 and 2,889 observations per commodity.

The accuracy statistic is the out-of-sample R^2 against the zero forecast,

$$R^2_{OS} = 1 - \frac{\sum_t (y_t - f_t)^2}{\sum_t y_t^2}, \qquad (3)$$

computed for each commodity. A positive value means a lower squared error than the zero forecast. We report the equal-weight mean across commodities (the panel mean R^2_{OS}) and the share of commodities with a positive value. We also report R^2_{OS} against the historical mean. The panel statistics use raw loss differences and loss differences divided by each commodity's mean squared return, so that high-variance commodities do not dominate. All 19 commodities are present on 2,869 of the 2,889 forecast dates (99.3%), and between 9 and 18 are present on the other 20. The panel mean R^2 is computed commodity by commodity and then averaged with equal weights, so missing dates do not change its weighting, while the panel loss tests average the loss difference over the commodities present on each date. Keeping only the dates with a full panel changes the panel mean R^2 by at most 0.13 percentage points for any model other than the naive forecast, and by at most 0.006 points for the shrunk models.

3.4. Statistical inference

Loss comparisons against the zero forecast. Let $d_{i,t} = y_{i,t}^2 - \left(y_{i,t} - f_{i,t}\right)^2$ be the loss difference in favor of the model, and let $\bar{d}_t$ be its average across commodities at date t. For non-nested models (the neural models and XGBoost) the primary test is a moving-block bootstrap of the mean of $\bar{d}_t$ under the null of zero mean [20]. We subtract the sample mean from the series $\bar{d}_t$ so that the null of zero mean holds in the resampled data, draw circular blocks of length $\max(20, 2h)$ with replacement until the original length is reached, and compute the mean of each resampled series, $\bar{d}_b^*$. The one-sided p-value is

$$p = \frac{1 + \#\{b: \bar{d}_b^* \geq \bar{d}\}}{B + 1}, \qquad (4)$$

with $B = 4999$ bootstrap resamples and a fixed random seed for each test. We also report the Diebold–Mariano statistic with the small-sample correction of Harvey, Leybourne and Newbold [21] and a HAC [22] t-test with lag $\max\left(h - 1, \lfloor 4(T/100)^{2/9} \rfloor\right)$. For $h = 5$ and $h = 21$, where targets overlap, we add HAC p-values with twice the baseline lag and the median p-value over the h non-overlapping subsamples formed by every h-th origin. Commodity-level p-values are adjusted with the procedure of Benjamini and Hochberg [23] within each model–horizon family.

Nested models. The zero forecast is the restricted model nested in the AR(1) and in the historical-mean forecast, whose coefficients are estimated by

expanding-window least squares, so for these two we use the Clark–West test as the primary test. Both are estimated by expanding-window least squares on observations whose targets are fully realized before the current month, and are re-estimated monthly: the AR(1) regresses the h-day target on the current return with an intercept, and the historical mean averages all earlier returns. Setting all estimated parameters to zero returns the zero forecast. Against a zero forecast the Clark–West adjusted differential is

$$y_t^2 - [(y_t - f_t)^2 - f_t^2] = 2y_t f_t, \qquad (5)$$

which tests whether the average product of the return and the forecast is positive. For non-nested models we report it only as a diagnostic. A positive Clark–West statistic together with a negative R^2_{OS} is not evidence of predictive content in the sense of lower squared error.

Simulation-based diagnostic of test size. To check the size of the tests we simulated 2,900 daily returns under two processes: i.i.d. Gaussian returns, and Student-$t(6)$ innovations with GARCH(1,1) volatility ($\omega = 0.05$, $\alpha = 0.08$, $\beta = 0.90$). Forecasts contained persistent AR(1) noise (coefficient 0.9), and targets overlapped for $h > 1$. There were 150 replications per case, and the bootstrap used 299 resamples. At the 5% level, rejection rates under the heavy-tailed process were 4.7% to 8.7% for the block bootstrap, 5.3% to 8.7% for Clark–West and 8.7% to 15.3% for the base-lag HAC loss test. With 150 replications, a rejection rate near 5% has a Monte Carlo standard error of about 1.8 percentage points and one near 9% about 2.3, and AR(1)-noise forecasts do not reproduce the loss-difference distribution of trained networks. The simulation is therefore a diagnostic. It motivates the bootstrap as the primary test for non-nested comparisons and does not establish the validity of any test in general.

Confidence bounds. To state what the data can rule out, we compute confidence intervals for the panel mean R^2_{OS}. We resample the same blocks of dates for every commodity (block length $\max(20, 2h)$, $B = 4999$), recompute each commodity's R^2_{OS} and average them. The one-sided 95% upper bound is the 95th percentile of the bootstrap distribution. This is a percentile interval, and we did not simulate its coverage for the forecasts studied here, so we interpret it as a description of what the sample leaves open and not as an exact guarantee.

3.5. Cross-commodity training experiments

These experiments use N-BEATS at $h = 1$. They compare the following training strategies, with the same seeds, re-estimation schedule, scaling and early stopping as in Section 3.2.

- **Local.** One model per commodity with its own tuned hyperparameters (the main-study model).
- **Global.** One model on all 19 commodities with hyperparameters tuned on the pooled data.
- **LocalHP.** One model per commodity, but with the hyperparameters of the global model. This comparison checks whether the advantage of the global model persists under a common hyperparameter configuration. It does not isolate a causal effect of pooling, because training-sample size and the fit of the hyperparameters to each method still differ.
- **Sector.** Three pooled models, one for energy (5 commodities), one for metals (gold, silver, platinum, palladium, copper) and one for agriculture and livestock (9 commodities, called agriculture below: corn, wheat, soybeans, coffee, sugar, cotton, cocoa, live cattle, lean hogs), all using the global hyperparameters.
- **Random partitions.** Twenty random partitions of the 19 commodities into groups of 5, 5 and 9, each trained like Sector. The partitions are drawn with fixed seeds (1000 to 1019).

Hypotheses. H1: the Global arm has a lower pooled normalized squared loss than the Local arm. We test it with the paired block bootstrap of Section 3.4, one-sided at the 5% level, and repeat it with LocalHP in place of Local. H2: the Sector arm has a higher panel mean R^2_{OS} than the random partitions. The p-value is

$$p = \frac{1 + \#\{r: R^2_{\text{rand},r} \geq R^2_{\text{Sector}}\}}{R + 1}, \qquad (6)$$

with $R = 20$ random partitions, so the smallest attainable value is $1/21 = 0.048$. For each sector we also compare the Sector arm with the random partitions on the same commodities. These are three separate tests without a multiple-testing adjustment, and we treat them as descriptive. We

also test the Sector arm against each random partition with a paired bootstrap. An arm beats the zero forecast only if its bootstrap p-value is below 0.05 and its mean R^2_{OS} is positive. The hypotheses were written down after the main study had been run, so all results of this section are exploratory. The main design was fixed in the configuration before the main forecast run (configuration hash cfe1464a1f84), and no public pre-registration was filed.

3.6. Forecast decomposition and calibration

Decomposition. For each commodity and arm define the cosine between the forecast and the target and the relative forecast amplitude,

$$c = \frac{\sum_t y_t f_t}{\sqrt{\sum_t y_t^2 \sum_t f_t^2}},$$

$$a = \sqrt{\frac{\sum_t f_t^2}{\sum_t y_t^2}}. \qquad (7)$$

Then

$$R^2_{OS} = 2ca - a^2 = c^2 - (a - c)^2. \qquad (8)$$

The same quantity can be written as a cross-product term minus a second-moment penalty, $R^2_{OS} = 2\sum yf/\sum y^2 - \sum f^2/\sum y^2$, and Equation (8) is this identity expressed through c and a. We report the cross-product form in Section 4.3 when we describe individual models and the c^2 form when we compare training strategies. The cross-product term can be negative, whereas c^2 cannot. The factor that minimizes squared error is $b^* = \sum yf/\sum f^2 = c/a$, so c^2 is the R^2_{OS} that the forecast would reach after the best possible rescaling without any restriction on its sign or size (the optimal factor is negative when $c < 0$), and $(a - c)^2$ is the loss from having the wrong amplitude. Because b^* is computed on the evaluation sample, c^2 describes the forecasts retrospectively and is not an implementable strategy; it is an optimistic measure of performance after rescaling and not an unbiased out-of-sample estimate, and it need not be attainable by the restricted procedure below, which limits the factor to [0,1]. We use it to compare arms and not to judge a level. We average each term across commodities with equal weights and test differences between arms by a moving-block bootstrap over dates (block length 20, 999 resamples, the same dates for all arms and commodities), reporting 95% percentile intervals. The decomposition is an accounting device and does not identify a mechanism.

Calibration with past data. To check whether differences between arms survive when amplitude is controlled, we rescale each arm's forecasts with the expanding-window procedure used for the shrunk forecasts in the main study. The rescaled forecast is $\tilde{f}_t = \hat{b}_t f_t$, where

$$\hat{b}_t = \min\left\{1, \max\left\{0, \frac{\sum_{s:\, s+h \le t} y_s f_s}{\sum_{s:\, s+h \le t} f_s^2}\right\}\right\}, \qquad (9)$$

and the sums run only over earlier origins whose targets are fully realized at t. We set $\hat{b}_t = 0$ until 250 such origins are available. The weight cannot reverse a forecast or increase it, so the procedure pulls forecasts toward zero. Some of the zero forecasts are mechanical: all forecasts before the 250-origin warm-up are zero by construction. The rescaled forecast is exactly zero at 48% to 70% of origins, depending on the arm. The warm-up accounts for 8.7% of origins in every arm, and the remaining zeros come from an estimated factor equal to zero. The high frequency of zero forecasts may limit the statistical power to distinguish the arms, so we repeat the comparisons on post-warm-up origins only. We repeat the tests of Section 3.5 on the rescaled forecasts.

Association with returns. For each arm we compute the Pearson and the Spearman correlation between the forecast and the target for each commodity and average them across commodities. Neither measure depends on forecast amplitude. We obtain 95% intervals with the same block bootstrap over dates; the Spearman ranks are computed on the full sample and held fixed in the bootstrap, which is an approximation that may understate the uncertainty in the ranks. These correlations use the raw forecasts.

3.7. Economic evaluation

Statistical accuracy and economic value are different questions, so we also evaluate the forecasts from the perspective of a mean–variance investor, following Campbell and Thompson [8]. At each origin the investor holds in commodity i the position

$$w_{i,t} = \min\left\{1.5,\ \max\left\{0,\ \frac{f_{i,t}}{\gamma\, \hat{\sigma}^2_{i,t}}\right\}\right\}, \qquad (10)$$

where $\hat{\sigma}^2_{i,t}$ is the sample variance of the previous 60 daily returns up to and including t and γ is the coefficient of relative risk aversion. The rest of the capital is held in cash, and futures returns are treated as excess returns. The net return over day $t+1$ is

$$R_{i,t+1} = w_{i,t}(e^{y_{i,t+1}} - 1) - c\,|w_{i,t} - w_{i,t-1}|, \qquad (11)$$

with a one-way transaction cost c of 0, 5, 10 or 20 basis points. The portfolio gives equal capital to the commodities present on each date. Its annualized certainty-equivalent return is

$$\mathrm{CER} = 252\left(\bar{R} - \frac{\gamma}{2}\,\widehat{\mathrm{Var}}(R)\right). \qquad (12)$$

The benchmarks are the same investor using the historical-mean forecast (HistMean), the equal-weight buy-and-hold portfolio with $w_{i,t} = 1$ (1/N), and cash. We test whether a strategy's CER exceeds that of a benchmark with a one-sided moving-block bootstrap over dates (blocks of 20 days, the same dates for both strategies). We also compute the break-even cost, the one-way cost at which a strategy's CER falls to that of the benchmark. The main setting is $\gamma = 3$; robustness checks use $\gamma = 5$, positions between −1 and 2, a 120-day variance window, the three subperiods, and the exclusion of days inside detected roll windows (Section 6). The returns come from the same continuous price series as the forecasts and are not roll-adjusted, so the exercise measures the value of the forecasts for a stylized investor and not the profit of an implementable futures strategy.

neural architectures. Every unshrunk neural model has a negative mean out-of-sample R^2 at h=1. The values for N-BEATS, N-BEATSx and N-HiTS are −3.95%, −5.32% and −3.35%. In none of these cases does the loss-based test come close to rejecting the zero forecast (bootstrap p-values of 1.00 for the first two and 1.00 for N-HiTS). The two variants of N-BEATSx with lagged macro inputs do worse still, at −6.77% and −7.27%. They were added after the main analysis and we report them for completeness only.

Training one network on all 19 commodities moves these numbers much closer to zero, to −0.63% for N-BEATS and −0.47% for N-HiTS, but not above it. Rescaling the forecasts with the past-only procedure of Section 3.6 brings the neural models even closer: the mean R^2 is 0.03% for shrunk N-BEATSx and between −0.05% and −0.02% for the other shrunk models in Table 1. The bootstrap p-values of these shrunk models run from 0.34 to 0.88, so none of them shows a gain over zero.

The one model that meets the Clark–West criterion in the configuration is the AR(1). Its mean R^2 is 0.02% and its panel Clark–West p-value is 0.003. The squared-error gain is nevertheless indistinguishable from zero: the block-bootstrap p-value is 0.15 with raw weights and 0.38 with variance-normalized weights, and the Diebold–Mariano p-values are 0.19 and 0.38. The smallest panel loss-test p-value for any model, horizon, test variant or weighting is 0.146. At h=5 and h=21 the neural models are further from zero, with mean R^2 of −8.6% and −18.4% at h=5 and −29.4% and −79.4% at h=21, and none of the panel Clark–West tests points to an improvement.

4. Results

4.1. Accuracy against the zero forecast

Table 1 gives the one-day results for the main models and the longer-horizon results for the two

Horizon	Model	Mean R^2 (%)	Assets with $R^2 > 0$	Panel CW p	Panel loss-test p
1	AR1	0.02	9/19	0.0030	0.380
1	N-BEATS	−3.95	0/19	0.7947	1.000
1	N-BEATSx	−5.32	1/19	0.9260	1.000
1	NBEATSxL (post hoc)	−6.77	0/19	0.9730	1.000
1	NBEATSx_1D (post hoc)	−7.27	0/19	0.9727	1.000
1	N-HiTS	−3.35	0/19	0.5051	1.000
1	N-BEATS (global)	−0.63	3/19	0.0725	1.000
1	N-HiTS (global)	−0.47	3/19	0.7533	0.996

Horizon	Model	Mean R^2 (%)	Assets with $R^2 > 0$	Panel CW p	Panel loss-test p
1	Shrunk N-BEATSx	0.03	5/19	0.0509	0.335
1	Shrunk NBEATSxL	−0.04	6/19	0.1636	0.780
1	Shrunk NBEATSx_1D	−0.02	5/19	0.0925	0.642
1	Shrunk N-BEATS (global)	−0.02	3/19	0.0516	0.684
1	Shrunk N-HiTS (global)	−0.05	2/19	0.4210	0.875
5	N-BEATS	−18.35	0/19	0.2662	1.000
5	N-HiTS	−8.61	0/19	0.2623	1.000
21	N-BEATS	−79.37	2/19	0.5406	1.000
21	N-HiTS	−29.37	1/19	0.8079	0.994

Table 1. Equal-weight panel mean out-of-sample R^2 against the zero forecast. R^2 is in percent of the variance of returns. Panel Clark–West p-values are one-sided with raw weights and serve as the primary test only for the nested AR(1); for the neural models they are a diagnostic. Panel loss-test p-values come from the block bootstrap with variance-normalized weights (4,999 resamples, fixed seed per test) and are the primary test for the neural models.

The two tests can disagree about a given commodity. Of 817 commodity–model–horizon results, 64 have a raw Clark–West p-value below 0.05, and 20 of those come with a negative R^2. Seventeen of the 20 belong to unshrunk neural models, of 23 such rejections for that group. A forecast that is too large and slightly correlated with the return can pass the covariance test while losing to zero on squared error. Gold at h=1 illustrates this. The Clark–West test flags N-BEATSx for Gold even though its R^2 is 0.01% and its loss-test p-value is 0.49. After adjusting for the false discovery rate within each model–horizon family, 12 Clark–West results remain, nine with a positive R^2. Among the non-nested models, no commodity–model–horizon combination survives the loss-based adjustment (the smallest q-value is 0.30). The only survivors under the primary-test rule are the AR(1) and the historical mean for Gold at h=21 (q=0.034 and 0.048, R^2 of 3.9% and 3.7%). Since the historical mean has almost the same R^2 as the AR(1), these results probably reflect the positive average return of gold over the sample and not information in lagged returns, although we have not tested that explanation directly.

Confidence bounds show what these tests leave open (Fig. 1). At h=1 the 95% intervals for the panel mean R^2 of all unshrunk neural models lie entirely below zero. For N-BEATS the interval runs from −4.5% to −3.3%, and for global N-BEATS from −0.9% to −0.3%. For the shrunk forecasts the one-sided 95% upper bounds lie between −0.03% and 0.16%. The largest is 0.16%, for shrunk N-BEATSx, and for the AR(1) it is 0.11%. Block lengths of 10 and 40 days instead of 20 do not change this: the largest upper bound over the three lengths for the shrunk models, the AR(1) and the historical mean is 0.156%. At h=21 the intervals are wider, and the data cannot rule out a mean R^2 of about 0.5% for the AR(1) or the historical mean. A gain of that size could matter at a monthly horizon [8], so the monthly results say less than the daily ones.

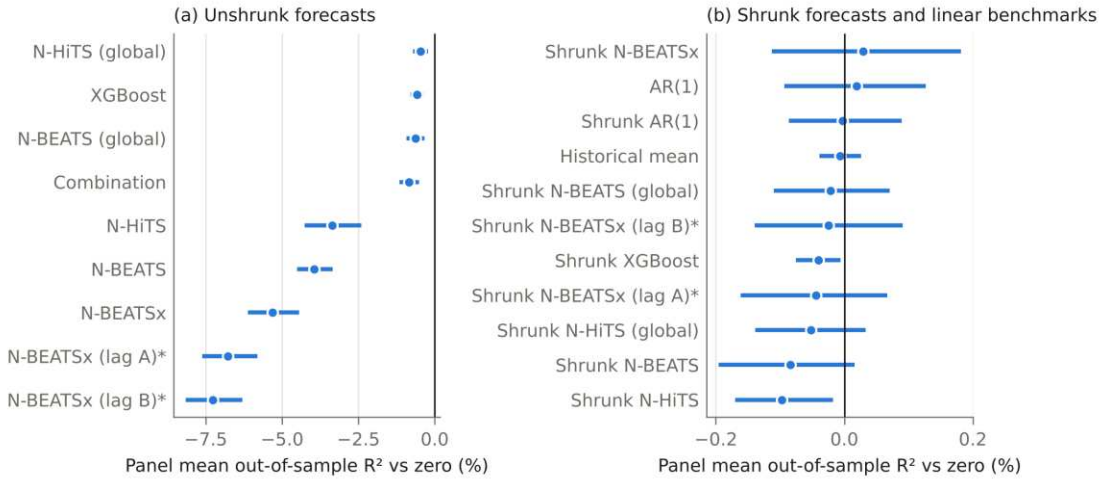


Fig. 1 *Panel mean out-of-sample R^2 against the zero forecast at h=1, in percent, with 95% percentile intervals from a moving-block bootstrap over dates (4,999 resamples, circular blocks of 20 trading days, the same resampled dates for all commodities); an interval that lies entirely below zero rules out a gain over the zero forecast at the 5% level. Panel (a) shows the unshrunk forecasts and panel (b) the shrunk forecasts and the linear benchmarks; the horizontal scales differ. Lag A uses macro inputs dated strictly before the commodity date and lag B shifts the standardized macro changes by one commodity trading day; both variants (*) were added after the main analysis*

4.2. Training strategies

Table 2 collects the results of the cross-commodity experiments. For loss the answer to H1 is yes. Global N-BEATS has a mean R^2 of −0.63%, against −3.95% for Local, and the paired bootstrap puts its normalized squared loss below that of Local (mean difference 0.0332, one-sided p=0.0002) and below that of LocalHP (0.1096, p=0.0002). LocalHP does much worse than Local, at −11.59%, so running the local models with the global hyperparameters lowers their accuracy considerably. The comparison with Local is the more natural one, since each arm has its own tuned configuration, and the comparison with LocalHP shows how sensitive the local models are to the configuration. Neither comparison isolates pooling, because the amount of training data and the fit of the hyperparameters differ between arms. None of the arms beats the zero forecast: the raw bootstrap p-values against zero are at least 0.98.

Arm	Raw R^2 (%)	Squared alignment c^2 (%)	Amplitude penalty (%)	Amplitude ratio	Shrunk R^2 (%)	Zero forecasts after shrinkage (%)
Local	−3.95	0.088	4.036	0.168	−0.084	57.1
LocalHP	−11.59	0.107	11.698	0.354	−0.115	48.5
Global	−0.63	0.079	0.706	0.087	−0.021	59.5
Sector	−2.11	0.145	2.255	0.142	−0.063	55.5
Random groups (mean of 20)	−1.52	0.129	1.650	0.125	−0.078	58.4

Table 2. Equal-weight panel means for N-BEATS at h=1. Raw R^2 equals c^2 minus the amplitude penalty (Equation 8). The amplitude ratio is a in Equation (7). The shrunk columns use the past-only rescaling of Section 3.6.

4.3. Forecast amplitude and signal

The decomposition of Section 3.6 shows where the loss differences come from. For the unshrunk N-BEATS models the standard deviation of the forecasts is 16.7% of the return standard deviation, and it is 22.1% for N-BEATSx and 23.2% for NBEATSxL. This is smaller than the standard deviation of returns, but 4.5 to 6.3 times that of the AR(1) forecast. For NBEATSxL the cross-product term of the identity in Section 3.6 is −0.40% and the second-moment penalty is 6.37% of the variance of returns, which gives the R^2 of −6.77% in Table 1. The forecasts are large relative to their weak covariance with realized returns.

Table 3 and Fig. 2 turn to the training strategies. The 3.32-point gap between Global and Local sits in the amplitude penalty, which is 3.33 points lower for Global, while the squared alignment term differs by −0.009 points and its interval, from −0.19 to 0.12, includes zero. The 10.96-point gap to LocalHP has the same shape (penalty −10.99, alignment −0.028 with an interval that includes zero). Pooling roughly halves the forecast amplitude, from 0.168 to 0.087. The alignment term itself is small for every arm, between 0.08% and 0.15% of the variance of returns. These intervals are about ±0.15 points wide and do not rule out small differences in alignment, so the data support the statement that pooled models produce smaller forecasts and do not show that they find more signal.

Contrast	Δ Raw R^2	Δ Alignment c^2	Δ Amplitude penalty
Global – Local	3.32 [2.65, 3.92]	−0.009 [−0.189, 0.119]	−3.33 [−3.99, −2.65]
Global – LocalHP	10.96 [9.16, 12.90]	−0.028 [−0.191, 0.078]	−10.99 [−12.92, −9.29]
LocalHP – Local	−7.64 [−9.71, −5.84]	0.019 [−0.124, 0.155]	7.66 [5.89, 9.67]
Sector – Global	−1.48 [−2.62, −0.46]	0.066 [−0.082, 0.332]	1.55 [0.56, 2.85]
Sector – Local	1.84 [0.52, 3.04]	0.057 [−0.171, 0.356]	−1.78 [−2.99, −0.30]
Global – Random groups	0.89 [0.57, 1.20]	−0.050 [−0.135, 0.026]	−0.94 [−1.29, −0.60]
Sector – Random groups	−0.59 [−1.49, 0.25]	0.016 [−0.121, 0.245]	0.61 [−0.15, 1.60]

Table 3. Differences between arms in percentage points, with 95% percentile intervals from a moving-block bootstrap over dates (999 resamples, circular blocks of 20 trading days, the same resampled dates for all commodities and arms, so that cross-commodity correlation is preserved). Δ Raw R^2 equals Δ alignment minus Δ amplitude penalty.

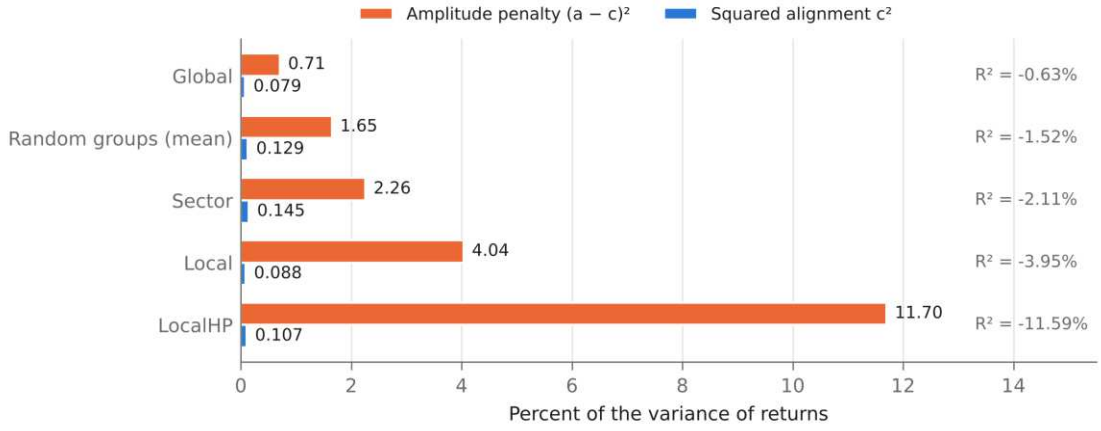


Fig. 2 *Decomposition of the panel mean out-of-sample R^2 of N-BEATS at h=1 for each training strategy, in percent of the variance of returns. R^2 equals the squared alignment c^2 minus the amplitude penalty $(a - c)^2$ (Equation 8). The random-group bar is the mean over 20 partitions*

The correlations between forecasts and returns do not follow the loss ranking either. Averaged over commodities, the Pearson correlation is −0.0001 for Local, 0.006 for Global, 0.012 for Sector and 0.019 for LocalHP, and only the interval for LocalHP excludes zero (0.002 to 0.035; Spearman 0.008, with an interval of 0.0001 to 0.017). LocalHP is the arm with the largest loss. The difference between LocalHP and Local in the Pearson correlation is 0.019 (0.002 to 0.034), while no Spearman or other Pearson contrast excludes zero. We did not adjust these intervals for the number of comparisons, and correlations of this size are tiny. They show that lower loss does not go with a stronger linear association between forecast and return.

Rescaling with past information gives a consistent picture. After the expanding-window rescaling the mean R^2 gap between Global and Local falls from 3.32 to 0.06 points (two-sided p=0.415), and the gap between Global and LocalHP falls from 10.96 to 0.09 points (p=0.182). No arm beats the zero forecast after rescaling, and every bootstrap p-value against zero is at least 0.62. The rescaling sets forecasts to zero at between 48% and 70% of origins, depending on the arm. The 250-origin warm-up explains only 8.7 percentage points of this, and the rest reflects an estimated weight of zero, that is, past forecasts whose covariance with returns was not positive. Restricting the comparison to post-warm-up origins leaves the picture unchanged: the gap between Global and Local is 0.07 points (two-sided p=0.39), the gap between Global and LocalHP is 0.10 points (p=0.13), the permutation p-value for Sector against random groups is 0.57, and no arm beats zero (bootstrap p-values of at least 0.69). A shrunk forecast that is zero much of the time still sits close to the benchmark whatever the network has learned, so we read the narrowing as consistent with the amplitude explanation and not as an independent confirmation of it.

4.4. Grouping by sector

H2 is not supported. In the raw forecasts the Sector arm has a mean R^2 of −2.11%, below each of the 20 random partitions, which range from −1.87% to −1.24% with a mean of −1.52%, so the permutation p-value for Sector beating random grouping is 1.00. The bootstrap interval for the difference from the random mean, −0.59 points, runs from −1.49 to 0.25, which includes zero, and its alignment and penalty components do not differ reliably (Table 3). Sector beats Local by 1.84 points (one-sided p=0.014), since pooling within sectors still shrinks the forecasts, and it falls 1.48 points behind Global.

The three sectors behave differently. In the raw forecasts the Sector arm earns −5.87% on the five energy contracts against −1.73% for random groups containing them, 4.14 points below (interval −7.36 to −1.24), and the whole difference is in the amplitude penalty; every random partition does better, so the permutation p-value is 1.00. For agriculture and livestock the Sector arm earns −0.73% against −1.57%, 0.84 points above (0.54 to 1.15), again through a lower penalty (−0.86) with no difference in alignment (−0.02, interval −0.12 to 0.08). No random partition does as well, which gives the smallest permutation p-value that 20 partitions allow, 0.048. For metals the figures are −0.83% against −1.21%, a difference of 0.39 points (−0.58 to 1.30; permutation p=0.14).

After past-only rescaling the Sector arm has a mean R^2 of −0.06%, and 11 of the 20 random partitions do at least as well (permutation p=0.57). The metals sector reaches 0.115% against −0.011% for random groups of the same size (permutation p=0.048), agriculture and livestock −0.015% against −0.047% (p=0.095), and energy −0.327% against −0.200% (p=0.86). The smallest attainable p-value turns up in agriculture for the raw forecasts and in metals for the rescaled ones, so the two versions of the forecast do not agree on which sector benefits. With three tests per version and no adjustment for multiple testing, neither result is evidence that sector grouping helps. These comparisons are exploratory.

4.5. Macro timing and subperiods

Lagging the macro inputs does not help. The two lagged variants have a mean R^2 of −6.8% and −7.3%, against −5.3% for the same-date N-BEATSx. The one case where a same-date result is significant after false-discovery adjustment under the Clark–West test, gold, disappears under lagging and does not survive the loss-based adjustment (q=0.30). Differences between same-date and lagged versions come from individual commodities and cannot be attributed to timing.

Subperiod results counsel caution. In 2020–2022 the mean R^2 is 0.19% for shrunk N-BEATSx and 0.10% and 0.11% for the shrunk lagged variants, with panel Clark–West p-values of 0.020, 0.018 and 0.019. In 2015–2019 and in 2023 to June 2026 the lagged variants are negative and not statistically

convincing. The subperiod tests use the Clark–West statistic without correction for the number of periods and models examined, so the 2020–2022 result is not evidence of stable predictability in a particular regime.

4.6. Economic value for a mean–variance investor

Table 4 reports the main economic results. Without transaction costs, several forecasts give the investor a higher CER than the historical-mean forecast. With $\gamma = 3$, the HistMean investor earns a CER of 1.38% a year. N-BEATSx raises this by 7.55 percentage points (p=0.003), the forecast combination by 4.31 points (p=0.009), the AR(1) by 2.74 points (p=0.031) and shrunk N-BEATSx by 2.07 points (p=0.018). A forecast with a negative out-of-sample R^2 can therefore carry economic value: the bounded position in Equation (10) removes most of the amplitude penalty that dominates squared-error loss, and what remains is the sign and rough size of the forecast. With $\gamma = 5$ the pattern is the same, and N-BEATSx gains 5.83 points (p=0.012).

Strategy	CER, 0 bp (%)	Gain vs HistMean, 0 bp (pp)	Gain vs 1/N, 0 bp (pp)	Gain vs HistMean, 5 bp (pp)	Break-even cost vs HistMean (bp)	Daily turnover
AR(1)	4.12	2.74 (0.031)	−2.65 (0.788)	−1.74 (0.865)	3.1	0.36
XGBoost	3.13	1.76 (0.074)	−3.63 (0.868)	−2.73 (0.990)	2.0	0.36
N-BEATS (Local)	5.29	3.91 (0.051)	−1.47 (0.725)	−4.01 (0.943)	2.5	0.64
N-HiTS	5.32	3.94 (0.052)	−1.44 (0.703)	−3.94 (0.947)	2.5	0.63
N-BEATSx	8.93	7.55 (0.003)	2.17 (0.237)	−1.20 (0.664)	4.3	0.70
N-BEATS (Global)	5.03	3.65 (0.061)	−1.73 (0.732)	−3.03 (0.878)	2.7	0.54
N-HiTS (Global)	3.45	2.07 (0.132)	−3.32 (0.847)	−4.70 (0.995)	1.5	0.54
Combination	5.68	4.31 (0.009)	−1.08 (0.631)	−2.79 (0.939)	3.0	0.57
N-BEATS (LocalHP)	7.56	6.18 (0.011)	0.79 (0.383)	−2.67 (0.839)	3.5	0.71
N-BEATS (Sector)	4.38	3.00 (0.128)	−2.38 (0.808)	−4.52 (0.950)	2.0	0.60
Shrunk N-BEATSx	3.45	2.07 (0.018)	−3.31 (0.781)	−0.33 (0.623)	4.3	0.20
Shrunk N-BEATS (Global)	1.42	0.04 (0.495)	−5.34 (0.900)	−1.15 (0.878)	0.2	0.10

Table 4. Economic value for a mean–variance investor with $\gamma = 3$, positions between 0 and 1.5 and a 60-day variance window, h=1. CER is annualized, in percent. Gains are differences in CER in percentage points per year, with one-sided p-values in parentheses from a moving-block bootstrap over dates (circular blocks of 20 trading days, the same dates for the strategy and the benchmark). The HistMean investor earns 1.38% without costs and 1.29% at 5 bp, and the 1/N buy-and-hold portfolio earns 6.76%. The break-even cost is the one-way cost at which the strategy's CER equals that of the HistMean investor. Daily turnover is the mean absolute change in the position per commodity. LocalHP and Sector are the arms of Section 3.5.

Figure 3 traces the certainty-equivalent returns across cost levels, and Fig. 4 sets them against the out-of-sample R^2. Two features limit this result. First, no strategy beats the 1/N buy-and-hold portfolio, which earns 6.76% a year without costs. N-BEATSx comes closest, 2.17 points above it, with

a p-value of 0.24. The HistMean investor holds small positions, so part of the gains over it reflects exposure to the positive average return of commodities over the sample, which the 1/N portfolio captures without any forecast. Second, the gains disappear once trading costs are charged. The raw neural forecasts change the position by 0.5 to 0.7 of capital per commodity each day, and the break-even cost against the HistMean investor lies between 1.5 and 4.3 basis points one way. Against 1/N only N-BEATSx, the variant with lagged macro inputs and LocalHP have any gain before costs, with break-even costs of 1.2, 0.2 and 0.4 basis points. At 5 basis points no strategy beats either benchmark. At 10 and 20 basis points every forecast-based strategy has a negative CER over the full sample, for both values of γ, both position limits and the 120-day variance window. In the subperiods at 10 basis points the only positive values are N-BEATSx and its one-day-lag variant in 2020–2022 (6.10% and 2.70%) and shrunk global N-BEATS in the same years (0.54%), all below the 12.31% of 1/N in that period and none significantly above the HistMean investor.

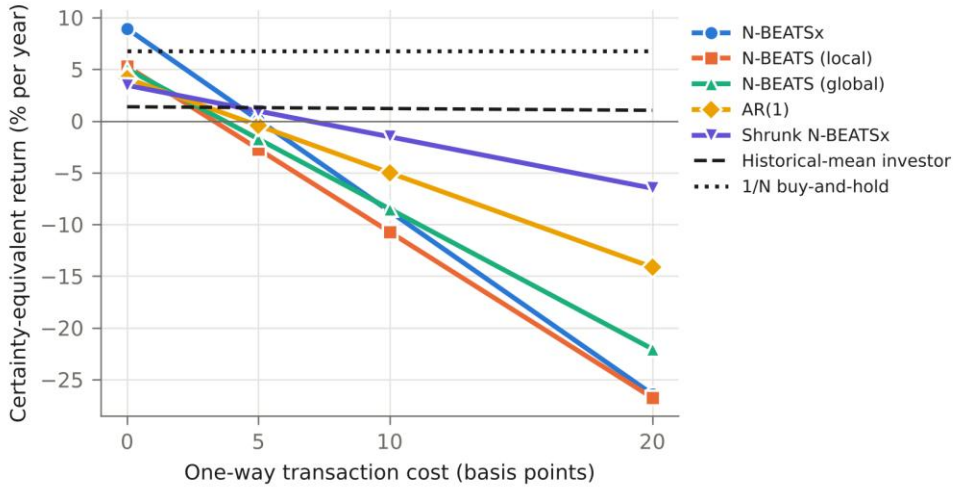


Fig. 3 *Certainty-equivalent return of the mean–variance investor (γ = 3, positions between 0 and 1.5) as a function of the one-way transaction cost, for selected forecasts and for the two benchmarks. The 1/N buy-and-hold portfolio does not trade, so its value does not depend on the cost*

The economic results also reverse the ranking of the training strategies. Without costs, LocalHP, the arm with the largest squared-error loss (R^2 of −11.59%), has the highest CER among the N-BEATS arms, 7.56%, against 5.29% for Local, 5.03% for Global and 4.38% for Sector. This is in line with Section 4.3, where LocalHP was the only arm whose correlation with returns was distinguishable from zero. In the main setting with costs of 10 basis points, the CER gain of Global over Local is 2.22 points (p=0.17) and that of shrunk Global over shrunk Local is 0.65 points (p=0.23), with both arms far below cash. The loss advantage of pooled training is thus not matched by an economic advantage.

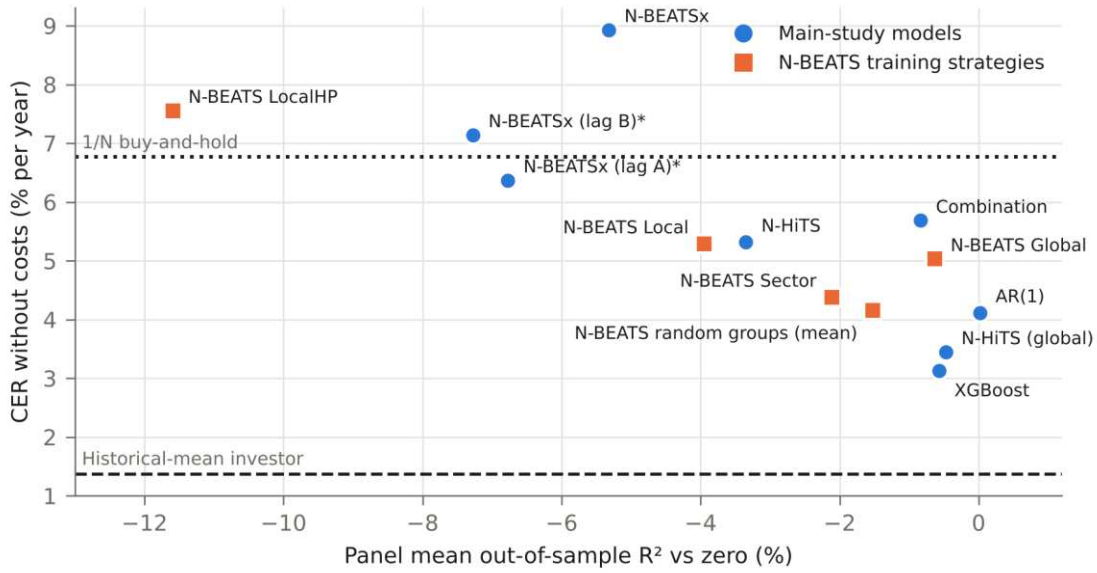


Fig. 4 *Out-of-sample R^2 against the zero forecast and certainty-equivalent return without transaction costs (γ = 3, positions between 0 and 1.5), at h=1. Circles are the main-study models and squares the N-BEATS training strategies of Section 3.5. Lag A and lag B (*) are the post-hoc variants with lagged macro inputs. Horizontal lines mark the historical-mean investor and the 1/N buy-and-hold portfolio*

Roll effects in the continuous series could in principle create apparent timing ability. When the returns of days inside detected roll windows (15% of days) are left out for the strategies and the benchmarks alike, the zero-cost gains of the strongest forecasts hardly change: 7.41 points for N-BEATSx (p=0.002), 4.49 points for the combination (p=0.010), 4.53 points for local N-BEATS (p=0.043) and 7.25 points for LocalHP (p=0.003). For Global N-BEATS, the Sector arm and XGBoost the gain falls below the range produced by dropping the same share of randomly chosen days, so part of their gain is associated with the detected windows. As Section 6 explains, the detector is a weak proxy for contract switches.

5. Discussion

The protocol separates three questions that a ranking of models by loss runs together. The first is whether the neural forecasts beat a forecast of zero. For these models and inputs they do not, although the confidence bound still leaves room for predictability that would matter to an investor. The second is what a relative gain consists of. The large loss gains from pooled training reflect smaller forecasts rather than more information about returns. The third is whether the forecasts are worth anything. They carry some directional information that a bounded investor can use before costs, but not enough to beat holding the commodities or to survive the costs of trading daily, which would have to stay below a few basis points one way, before roll costs, for any gain to remain.

Angelidis et al. [3] report a related problem for monthly linear forecasts, whose time-series versions are too variable. Our results show that forecast scale also dominates comparisons of neural networks at the daily horizon, but we cannot establish that the mechanism is the same. Pooled estimation may dampen commodity-specific forecasts, or the pooled hyperparameters may regularize more; the poor performance of LocalHP shows that the hyperparameters matter, and the experiments do not separate the two explanations. The reversal between the loss ranking and the certainty-equivalent ranking follows from the same arithmetic. Squared error penalizes the size of a forecast heavily, while a bounded position uses mainly its sign, so a training strategy can rank last on one criterion and first on the other.

Storage theory and hedging-pressure theories tie returns to forces that differ across energy, metals and agriculture, so networks trained within a sector might be expected to learn something that a network trained across sectors does not. The data do not show that: sector models did worse than random groupings in the raw forecasts, and the

sector-level differences ran in opposite directions for energy and agriculture, through forecast size rather than alignment. The theories may still matter for slower signals such as the basis or momentum, which the networks here do not see.

Two further points concern interpretation. A positive Clark–West statistic can coexist with higher squared error than zero, as it does for 20 of the 64 nominal rejections, so for non-nested forecasts the loss-based test is the relevant one. And our results need not conflict with those of Wang and Zhang [12], whose monthly models use richer commodity-specific predictors and are compared with AR(1) and forecast combinations rather than with zero.

6. Limitations

The data are the main limitation. Yahoo Finance continuous front-month prices are spliced from individual contracts at dates the source does not document, so some returns include the gap between contracts, and the source does not keep prices of expired contracts from which the switch dates could be reconstructed. All models share the same targets, so comparisons between models are less exposed to errors common to all of them than absolute levels are. They are not immune, because models may react differently to jumps at contract rolls. We ran two partial checks. Dropping the forecast origins whose targets lie within two trading days of a large jump in volume, a proxy for contract switches, removes between 5% and 19% of origins depending on the commodity and changes the panel mean R^2 by at most 0.45 percentage points. Every unshrunk neural model stays below zero, and the smallest loss-based panel p-value is 0.18. For N-BEATS the gain from dropping the flagged origins, from −3.95% to −3.50%, is larger than the gain from dropping the same share of randomly chosen origins (5th to 95th percentile of −4.22% to −3.80%), so roll-related noise appears to affect the raw forecasts, although it does not change the conclusion. The proxy is itself weak. In a permutation test, the flagged days carry larger returns than randomly chosen days at the 5% level for only five of the 19 commodities (natural gas, corn, soybeans, sugar and cotton), so the flagged days cannot be taken to be contract switches. For WTI, natural gas, heating oil and gasoline we also re-estimated N-BEATS and N-HiTS on roll-clean returns built from EIA settlement prices, which are available for these four contracts until April 2024, and they too fail to beat zero: mean R^2 is negative on both data sets, and only one of the eight contract–model pairs is positive on the roll-clean data (N-BEATS for natural gas, 0.3%). Neither check covers all 19 series, and contract-level data would be needed to settle the matter. The WTI observation of April 2020 was removed because its price was not positive, and the return across the gap spans more than one day.

The information set is narrow. The networks see past returns, and N-BEATSx sees the dollar index, the VIX and the ten-year Treasury yield. They do not see the basis, inventories, open interest or term structure, which carry the commodity-specific signals that the literature finds useful. Our negative result concerns this information set, and it says nothing about whether a model with richer inputs could do better. The macro inputs may also be timed differently from the commodity closes. The two lagged variants address the calendar-date side of this concern, but they do not reconstruct release times, and they were added after the main analysis.

Part of the inference was added after the first results were seen, and the cross-commodity experiments were designed after the main study. Their results are exploratory. The size simulations are a diagnostic: with 150 replications each rejection rate carries a Monte Carlo error of about two percentage points, and the forecasts in the simulation do not reproduce the loss differences of trained networks. The three sector comparisons are not adjusted for multiple testing, and with 20 random partitions the smallest attainable permutation p-value is 0.048. The decomposition term that measures alignment is computed on the evaluation sample and is not an implementable strategy. The past-only rescaling sets forecasts to zero at 48% to 70% of origins, of which the warm-up explains 8.7 percentage points and the rest reflects estimated weights of zero, so the close agreement of arms after rescaling may limit the statistical power to separate them. The false-discovery adjustment is applied within each model–horizon family and not across all specifications. At the 21-day horizon the bounds are wide, and the results there are less informative than at the daily horizon. Finally, the economic evaluation uses returns of the continuous price series, a stylized investor, proportional costs and no margin, roll, liquidity or execution data, so it measures the value of the forecasts and not the profit of an implementable futures strategy. Many strategies and settings were evaluated, and only the zero-cost

comparisons against the historical-mean investor are significant.

7. Conclusion

For the models and horizons studied, we find no significant panel improvement over a zero-return forecast in the loss-based tests. The one-sided upper bound on panel-mean daily R^2 is 0.16%. AR(1) meets the Clark–West criterion, with mean R^2 of 0.02%, but its estimated squared-error improvement is not statistically distinguishable from zero.

There are nevertheless substantial differences among neural training strategies. Global training lowers squared-error loss relative to separate training, but most of the raw difference lies in forecast amplitude rather than a detectable improvement in squared alignment. Sector grouping provides no reliable advantage over random partitions. In economic terms, several forecasts help a mean–variance investor relative to a historical-mean forecast before costs, but none beats equal-weight buy-and-hold, the gains vanish at one-way costs of a few basis points, and the training strategy with the lowest loss is not the one with the highest certainty-equivalent return. The results illustrate why ranking forecasting models is not the same as establishing return predictability: the zero forecast remains the relevant benchmark, and the decomposition helps identify what drives a relative loss reduction.

Two extensions would test these findings most directly. The first is contract-level data with a documented roll rule, which would settle the questions about the price series and allow a proper economic test. The second is an information set that includes the basis, momentum and inventories, which would show whether networks gain from slower, commodity-specific signals once their forecast amplitude is controlled.

Statements and Declarations

Funding. No funding was received for conducting this study.

Competing interests. The author has no relevant financial or non-financial interests to disclose.

Ethics approval and consent. Not applicable; the study uses publicly available market data and involves no human or animal participants.

Data availability. The futures and macroeconomic price series were retrieved from Yahoo Finance, and the settlement prices of NYMEX energy contracts 1 and 2 are from the U.S. Energy Information Administration. The stored out-of-sample forecasts are available from the author on reasonable request.

Code availability. The code and configuration files used in the study are available from the author on reasonable request.

References

1. Gorton, G., Rouwenhorst, K.G.: Facts and fantasies about commodity futures. Financial Analysts Journal 62(2), 47–68 (2006)

2. Moskowitz, T.J., Ooi, Y.H., Pedersen, L.H.: Time series momentum. Journal of Financial Economics 104(2), 228–250 (2012)

3. Angelidis, T., Sakkas, A., Tessaromatis, N.: Predicting commodity returns: Time series vs. cross sectional prediction models. Journal of Commodity Markets 38 (2025)

4. Fama, E.F., French, K.R.: Commodity futures prices: Some evidence on forecast power, premiums, and the theory of storage. Journal of Business 60(1), 55–73 (1987)

5. Bessembinder, H.: Systematic risk, hedging pressure, and risk premiums in futures markets. Review of Financial Studies 5(4), 637–667 (1992)

6. Hong, H., Yogo, M.: What does futures market interest tell us about the macroeconomy and asset prices? Journal of Financial Economics 105(3), 473–490 (2012)

7. Montero-Manso, P., Hyndman, R.J.: Principles and algorithms for forecasting groups of time series: Locality and globality. International Journal of Forecasting 37(4), 1632–1653 (2021)

8. Campbell, J.Y., Thompson, S.B.: Predicting excess stock returns out of sample: Can anything beat the historical average? Review of Financial Studies 21(4), 1509–1531 (2008)

9. Rapach, D.E., Strauss, J.K., Zhou, G.: Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. Review of Financial Studies 23(2), 821–862 (2010)

10. Cotter, J., Eyiah-Donkor, E., Potì, V.: Commodity futures return predictability and intertemporal asset pricing. Journal of Commodity Markets 31 (2023)

11. Merton, R.C.: An intertemporal capital asset pricing model. Econometrica 41(5), 867–887 (1973)

12. Wang, S., Zhang, T.: Predictability of commodity futures returns with machine learning models. Journal of Futures Markets 44(2), 302–322 (2024). https://doi.org/10.1002/fut.22471

13. Oreshkin, B.N., Carpov, D., Chapados, N., Bengio, Y.: N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In: International Conference on Learning Representations (2020)

14. Challu, C., Olivares, K.G., Oreshkin, B.N., Garza, F., Mergenthaler-Canseco, M., Dubrawski, A.: N-HiTS: Neural hierarchical interpolation for time series forecasting. In: Proceedings of the AAAI Conference on Artificial Intelligence 37(6), 6989–6997 (2023)

15. Olivares, K.G., Challu, C., Marcjasz, G., Weron, R., Dubrawski, A.: Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx. International Journal of Forecasting 39(2), 884–900 (2023)

16. Gu, S., Kelly, B., Xiu, D.: Empirical asset pricing via machine learning. Review of Financial Studies 33(5), 2223–2273 (2020)

17. Welch, I., Goyal, A.: A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21(4), 1455–1508 (2008)

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

19. Clark, T.E., West, K.D.: Approximately normal tests for equal predictive accuracy in nested models. Journal of Econometrics 138(1), 291–311 (2007)

20. Künsch, H.R.: The jackknife and the bootstrap for general stationary observations. Annals of Statistics 17(3), 1217–1241 (1989)

21. Harvey, D., Leybourne, S., Newbold, P.: Testing the equality of prediction mean squared errors. International Journal of Forecasting 13(2), 281–291 (1997)

22. Newey, W.K., West, K.D.: A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55(3), 703–708 (1987)

23. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B 57(1), 289–300 (1995)