Tail Calibration in Panel Volatility Forecasting: Evidence from Metropolitan Housing Markets

Vasilios Plakandaras[1]

Department of Economics

Democritus University of Thrace

Abstract

We forecast the conditional distribution of realized volatility for ten U.S. metropolitan housing futures markets over 2007–2025 using a panel Bayesian neural network with a regularized-horseshoe prior and a city-specific score-driven exponential generalized autoregressive conditional heteroskedasticity variance equation, benchmarked against thirteen classical panel specifications and Bayesian additive regression trees. Our findings verify the weak predictability of the second moment, as no specification improves on a pooled autoregression at any horizon. Against that bound, we document a result the volatility literature has not reported: the training objective, and not the choice of model, governs the direction in which a predictive density fails. Estimating the same network under a quasi-likelihood loss yields conservative densities with coverage above nominal, while estimating it under a likelihood yields sharper densities with overconfident tails across all horizons. Comparison with a tree ensemble carrying the same features but a fixed within-window variance identifies the mechanism, as the score-driven variance equation buys no average accuracy and instead serves as the channel through which the estimation loss reaches the tail. The consequence is economic rather than statistical. Under the Basel traffic-light test the two otherwise identical models occupy different capital zones under stress, so a decision internal to a modelling team and invisible to a supervisor reading the forecasts changes the capital a regulated holder must hold.

JEL codes: C53, C33, G17, R31, C45

Keywords*:* Realized volatility; panel forecasting; predictive density calibration; Bayesian neural networks; score-driven models; housing markets

Acknowledgment*:* Claude (Anthropic) was used to assist with data analysis code, figure preparation and manuscript editing. All results were verified by the author, who takes full responsibility for the content.

---

[1] University Campus, Komotini, Greece

vplakand@econ.duth.gr, +306944503768

1. Introduction

Consider two forecasting models no less accurate on average, estimated on the same data, differing only in the loss function minimized during training. One passes a supervisory back-test of its value-at-risk forecasts and leaves a bank's regulatory capital on its floor; the other, in places more accurate than the first, fails that back-test in a crisis and attracts a capital surcharge. The choice between them is internal to the modelling team and invisible to a supervisor reading only the forecasts, yet it moves required capital on an identical book of exposures. This paper documents that mechanism for the volatility of metropolitan housing markets, and in doing so reaches a broader and more sobering conclusion: for the second moment, the training objective governs how far a model can be trusted at the tail far more than the choice of how accurately the model forecasts, since accuracy itself is sharply bounded.

Housing price volatility is a direct input into mortgage default pricing, the hedging of mortgage-backed security portfolios and macroprudential surveillance (Miller and Peng, 2006; Fairchild, Ma and Wu, 2015; Cheng and Chiu, 2020). What all these applications require is not a point forecast but a calibrated predictive density, since capital charges, value-at-risk limits and stress thresholds are all functions of the tails rather than of the conditional mean. Forecasting density is also harder than forecasting the price level itself, for reasons idiosyncratic to the second moment of the distribution. Volatility is never observed: each month's value is itself an estimate formed from the dispersion of daily settlement returns within the month, so the target carries measurement error the price level rarely does, and that error is larger in precisely the turbulent, thinly traded months whose tails the forecast must capture. Short estimation windows compound this, identifying a conditional variance and the shape of its tail far more loosely than a conditional mean (Karoglou, Morley and Thomas, 2013; Plakandaras et al., 2020). Furthermore, the transmission of macro-financial shocks into the housing market is state-dependent, with shocks arriving in recessions producing deeper and more protracted responses than identical shocks arriving in expansions (Cheng and Chiu, 2020; Aastveit, Albuquerque and Anundsen, 2023). Cities are also not independent. A national component accounts for a substantial and time-varying share of metropolitan housing dynamics (Del Negro and Otrok, 2007; Tidwell et al., 2023), and its comovement structure differs across the price distribution rather than being constant (Nazlioglu et al., 2026), so that a single-market model discards information by construction. In tandem, credit supply is among the most sharply identified of these common drivers, explaining roughly half of the boom-era rise in price-rent ratios (Greenwald and Guren, 2025; Favara and Imbs, 2015; Justiniano, Primiceri and Tambalotti, 2019), with the evolution of credit supply being heterogeneous across markets (Loutskina and Strahan, 2015; Di Maggio and Kermani, 2017; Mian and Sufi, 2022).

Taken together, these features define the estimation problem. A forecasting model for metropolitan housing volatility must pool information across a small cross-section without imposing homogeneity, admit nonlinear and possibly unstable links to macro-financial conditions evolving with time, allow the conditional variance to be city-specific, time-varying and asymmetric, and deliver a predictive density whose tails can be evaluated rather than assumed.

The demand for such a framework is not merely academic. A volatility forecast for housing is an input to a chain of decisions that are, increasingly, made under formal evaluation regimes. Mortgage insurers and the housing-finance institutions price and reserve against the tail of a house-price loss distribution, while the capital held against that exposure is set, under the FHFA's Enterprise Regulatory Capital Framework, by a rule that responds to house-price deviations from trend (Federal Housing Finance Agency, 2020). In parallel, credit risk transferred to private investors through the STACR and CAS programmes is tranched against the same tail, on the premise that the reference pools are geographically diversified (Federal Housing Finance Agency, 2024). A regulated holder's internal risk model is itself validated by a supervisory backtest that counts how often realised losses breach the model's stated quantile (Basel Committee on Banking Supervision, 1996). Each of these is a statement about a predictive distribution — its tail, its calibration, its behaviour under stress — rather than about a point forecast, and each is evaluated at the level of individual metropolitan markets rather than the nation. A forecasting framework intended to serve them must therefore deliver not just an accurate central path but a calibrated, city-specific predictive density whose tails can be audited and must be honest about how far that density can be trusted.

Three literature strands that could supply such a framework solve only part of the problem. The realized-volatility literature provides a mature measurement and forecasting apparatus, from the case of evaluating volatility models against a realized rather than a squared-return proxy (Andersen and Bollerslev, 1998; Barndorff-Nielsen and Shephard, 2002; Andersen et al., 2003), to the heterogeneous autoregressive benchmark (Corsi, 2009), its asymmetric competitors (Nelson, 1991; Hansen and Lunde, 2005) and the recognition that the realized proxy is itself measured with time-varying error (Bollerslev, Patton and Quaedvlieg, 2016). That apparatus is built for a single, typically long, series and has no mechanism for borrowing strength across markets. So, applications to the housing market have so far been used to construct and model high-frequency indices rather than to forecast their conditional variance across markets (Bollerslev, Patton and Wang, 2016).

The panel forecasting literature supplies the pooling mechanism, through mean-group estimation under parameter heterogeneity (Pesaran and Smith, 1995), common correlated effects estimators that absorb unobserved common factors by cross-sectional averaging (Pesaran, 2006; Chudik and Pesaran, 2015; Aquaro, Bailey and Pesaran, 2021), penalised

panel estimation (Camehl, 2023), and Bayesian shrinkage designed for panel density rather than point forecasts (Liu, Moon and Schorfheide, 2020). The central finding, established for a panel of U.S. metropolitan house prices, is that no single approach dominates uniformly, because the return to pooling depends on the degree and correlation structure of parameter heterogeneity and on the time-series length available per unit (Pesaran, Pick and Timmermann, 2026). That evidence concerns price levels, is obtained within linear conditional-mean specifications, and leaves the conditional variance unmodelled. A recent literature does estimate the conditional mean and variance jointly inside a neural architecture and evaluates predictive densities (Goulet Coulombe, Frenette and Klieber, 2026), but for a single series, using a penalized maximum likelihood, and with a variance process common to the series forecast.

A third strand delivers the evaluation apparatus this paper depends on, though it has not been brought to bear on a panel of volatility forecasts. The theory of proper scoring rules establishes when a forecast comparison is meaningful (Gneiting and Raftery, 2007), and its extension to elicitability shows that the ranking of risk-measurement procedures depends on the scoring function used to compare them, with direct implications for how a supervisor should design a backtest and what incentives the design creates (Nolde and Ziegel, 2017). Because average scores can conceal poor performance precisely in the region a risk manager cares about, this literature has developed evaluation tools focused on the tail rather than the center, through scoring rules that weight the region of interest (Amisano and Giacomini, 2007; Diks, Panchenko and van Dijk, 2011), through censoring of the density outside it (Mitchell and Weale, 2023), and most recently through a formal notion of tail calibration for probabilistic forecasts (Allen et al., 2025). The same literature warns that scoring-rule comparisons are themselves sensitive to the scale of the outcome and to measurement error in it (Kleen, 2024), which matters here because realized volatility is a noisy proxy for the quantity being forecasted. We take that apparatus as given and asks a question it raises but does not answer: whether the loss used to estimate a model, as distinct from the score used to evaluate it, determines the direction in which its tail is miscalibrated. Answering it requires a setting in which the same model can be estimated under competing objectives and its density read market by market, which is what the panel of metropolitan housing volatility provides and what a framework that is simultaneously nonlinear, heteroskedastic at the city level and evaluated as a density allows us to build.

This paper closes explicitly that gap and makes four key contributions. First, building on the Bayesian neural network of Hauzenberger et al. (2025), we develop a panel formulation for realized-volatility forecasting in which nested feature sets compete for inclusion under a single global–local shrinkage mechanism (Carvalho, Polson and Scott, 2010; Piironen and Vehtari, 2017) rather than through a separate specification search. Second, and centrally, we replace the stochastic-volatility error of the aforementioned

architecture with a city-specific, score-driven exponential GARCH variance equation, which updates each city's variance state using the scaled score of its own predictive log-likelihood (Creal, Koopman and Lucas, 2013; Harvey, 2013; Umlandt, 2023) and whose filtered path converges to the Kullback–Leibler-optimal time-varying parameter path even under severe misspecification (Beutner, Lin and Lucas, 2026). On a third level, we estimate every specification under two training regimes — a quasi-likelihood (QLIKE) scheme robust to an imperfect volatility proxy (Patton, 2011) or a Gaussian/Student-t likelihood scheme — and compare them directly rather than adopting one by default that is the common practice in the field. The comparison identifies a sharpness-versus-calibration trade-off rather than dominance of either, tracing its source to the differing intensity of shrinkage the two objectives induce. Finally, we benchmark against the full classical panel toolkit — pooled, mean-group, pooled mean-group, common correlated effects and dynamic common correlated effects HAR variants and a panel HAR-GARCH — and, crucially, against Bayesian additive regression trees (BART), a nonlinear competitor with an honest posterior predictive density (Chipman, George and McCulloch, 2010). We evaluate the models not only on accuracy — average log predictive likelihood, quantile scores, the Diebold–Mariano (1995) test and the model confidence set (Hansen, Lunde and Nason, 2011) — but on the calibration of their predictive tails, where the training objective, rather than the choice of model, turns out to govern the outcome.

The design follows from the data: ten metropolitan markets observed monthly over 218 periods, a short-T, small-N panel in which unregularized networks overfit and hand-specified nonlinear models demand a specification search the sample cannot support. Estimation is by direct projection at each of the one-, six- and twelve-month horizons on a rolling sixty-month window, chosen to accommodate the business-cycle instability documented for U.S. house prices (Karoglou, Morley and Thomas, 2013; Plakandaras et al., 2020), and the feature sets are ordered by economic content — autoregressive terms, then realized shocks, then the macroeconomic state and its uncertainty — so the shrinkage prior faces a well-posed question at each step.

Our findings are about volatility itself, not a modelling technique, and three of them reorient how this forecasting problem should be approached. The first is a calibration result we believe to be new: the training objective governs the *direction* in which a volatility model's density errs — quasi-likelihood estimation producing conservative tails and likelihood estimation overconfident ones, from the same model on the same data — and because supervisory capital is set by counting tail breaches, this single choice moves a model between the green and yellow zones of the Basel traffic-light test, a material economic consequence of an ostensibly technical decision. It is here that our model earns its place: trained under the quasi-likelihood objective it holds its ninety-nine per cent tail inside the supervisory green zone at every horizon in both calm and crisis periods, which

no classical panel estimator and no tree ensemble matches, precisely the property the risk applications above demand. The second finding isolates why.

By setting the BNN against BART, which shares its nonlinear mean treatment but not its score-driven variance, we locate the BNN's contribution squarely in the variance process and show that it is this component, not the conditional mean, that makes the tail controllable. The third is a bound the literature has not previously drawn: the predictability of the *level* of metropolitan housing volatility is weak at every horizon, so the value of a sophisticated model here lies not in sharper point forecasts but in the calibration and supervisory robustness of the density it delivers — a deliberate shift of emphasis from accuracy to calibration that we argue these applications require. A cross-sectional result completes the picture: differences in forecastability across markets are small and align weakly with the level of local volatility rather than with the supply-elasticity characteristics the urban-economics literature emphasises (Saiz, 2010; Aastveit, Albuquerque and Anundsen, 2023), a caution for risk-transfer pools diversified on geography. Underpinning all four is a methodological result: network width and depth matter far less than the information set and the training regime, so it is the shrinkage prior rather than any architecture search that governs performance. The paper is structured as follows: Section 2 describes the panel and the model, Section 3 reports the results, and Section 4 concludes, with Appendix A reporting cross-city and full-model detail and Appendix B the computational specifics.

2. Methodology and the Data

2.1 The Data

The panel is built from the settlement prices of the housing futures contracts listed on the Chicago Mercantile Exchange[2], which are written on the S&P CoreLogic Case-Shiller home price indices for individual metropolitan areas. Ten contracts are traded — Boston, Chicago, Washington DC, Denver, Las Vegas, Los Angeles, Miami, New York, San Diego and San Francisco. The sample runs monthly from September 2007 to October 2025, giving 218 months for each of the 10 cities and 2,180 city-month observations in total, with the cross-section determined by data availability. The ten contracts span the range of rental segmentation and financial integration documented by Greenwald and Guren (2025) and Loutskina and Strahan (2015), running from rate-sensitive coastal markets such as San Francisco and New York to less integrated interior markets such as Denver and Las Vegas, and they are the same markets that dominate the national-versus-

[2] Series identifier CME-HOUSING CITY CONT. WVOL - SETT. PRICE, continuous volume-weighted contract, settlement prices. Contract specifications are available at https://www.cmegroup.com/markets/real-estate.html.

local decompositions of Del Negro and Otrok (2007) and Gupta et al. (2023), making the pooling-versus-heterogeneity question studied here interesting.

Using traded contracts rather than the underlying indices is a substantive choice. A repeat-sales index is a smoothed, revised statistic published with a lag of roughly two months, and volatility computed from it is contaminated by smoothing rather than reflecting the price uncertainty faced by market participants. Futures settlement prices are observed daily, are not revised, and represent positions taken in real time. They also correspond directly to risk-management applications, as an institution hedging a portfolio of mortgage exposures does so, if at all, in these contracts, and the volatility that matters to it is the volatility of the instrument it can trade.

As already mentioned, the dependent variable is monthly realized volatility, in logarithmic form. For each city, realized volatility in month t is the sum of squared daily log returns on the settlement price within that month, which is the standard nonparametric estimator of the quadratic variation of the underlying price process (Barndorff-Nielsen and Shephard, 2002; Andersen et al., 2003). In 386 city-month observations the realized volatility is a structural zero and these are months in which the contract did not trade, and the settlement price was carried forward unchanged, so that every daily return within the month is zero. These months are spread across the panel rather than concentrated in a few thin markets, ranging from 25 observations in San Francisco to 52 in New York, between eleven and twenty-four per cent of each city's sample, and they are distributed evenly over the sample period. Given that they are a feature of a thinly traded market rather than missing data they are forward filled from the most recent non-zero value within the same city before logs are taken. This is a conservative treatment, assuming that an untraded month carries the volatility of the most recent traded one, rather than imputing calm that was never observed[3].

Table 1 reports the resulting series. Pooled across all observations, log realized volatility averages −9.67 with a standard deviation of 2.07 and a range from −15.45 to −3.29, spanning roughly five orders of magnitude in the underlying level, justifying the selection of logarithmic transformation of the series. City means are tightly grouped, from −9.97 in Washington DC to −9.18 in San Francisco, but that similarity in level conceals differences in dispersion and tail behaviour, which are already visible in the higher moments reported in the last two columns. The within-month return distribution is mildly

[3] Forward-filled months constitute 16.9 per cent of the evaluation sample. Excluding them leaves every calibration result intact: coverage under quasi-likelihood training moves from 0.992, 0.984 and 0.979 to 0.991, 0.984 and 0.980 at the one-, six- and twelve-month horizons, and under least-squares training from 0.925, 0.888 and 0.839 to 0.915, 0.887 and 0.839, so the gap between the objectives is 0.076, 0.097 and 0.141 against 0.068, 0.096 and 0.140 on the full sample. The exclusion widens the gap slightly at the short horizon, since the carried-forward months are the calmest in the sample and their removal worsens the least-squares under-coverage rather than the quasi-likelihood over-coverage.

right-skewed in every city, from near-symmetry in Washington DC (0.13) to pronounced asymmetry in Miami and Los Angeles (0.91–0.94), and uniformly heavy-tailed, with realized kurtosis clustered tightly around fourteen to fifteen. The near-constant excess kurtosis is what motivates the Student-t predictive density adopted below, while the cross-city spread in skewness foreshadows the heterogeneity in tail forecastability documented later in Section 3.

Table 1. Descriptive statistics: log realized volatility by city

City	Mean	Std. dev.	Min	Max	Skewness	Kurtosis
Boston	-9.957	2.019	-14.871	-5.774	0.403	15.406
Chicago	-9.407	1.989	-13.941	-5.121	0.267	14.237
Washington DC	-9.971	2.153	-14.860	-5.348	0.129	14.644
Denver	-9.636	1.932	-14.767	-4.725	0.906	14.768
Las Vegas	-9.560	2.185	-14.669	-3.294	0.793	14.149
Los Angeles	-9.767	2.159	-15.298	-4.812	0.914	14.518
Miami	-9.768	2.062	-15.303	-4.942	0.945	14.190
New York	-9.834	1.905	-14.222	-4.521	0.467	14.693
San Diego	-9.632	2.143	-15.445	-4.908	0.796	14.501
San Francisco	-9.183	2.055	-14.744	-4.592	0.543	14.802
Pooled	-9.667	2.072	-15.445	-3.294	0.616	14.591

Note: September 2007 to October 2025, computed after forward filling the 386 structural zeros described in the text. The pooled row covers all 2,180 city-month observations. Skewness and Kurtosis are the sample mean of the within-month realized skewness and realized kurtosis of daily settlement-price returns.

Four information sets are compared, each built by adding a block of predictors to the one before it. The purpose of organizing the predictors this way is to make the question "does this information earn its place?" answerable one block at a time, under a single estimation procedure, rather than through a sequence of separate specification tests.

The base set consists of twelve-monthly lags of a city's own log realized volatility, as it spans a complete seasonal cycle, yielding an autoregressive approach. Housing transaction volume, and with it realized volatility, rises predictably in the spring buying season and falls in winter, so a lag structure that stops short of twelve months forces the model to absorb a deterministic seasonal pattern into its residual, while a structure extending well beyond twelve adds parameters that duplicate information the seasonal lags already carry. We make two significant distinctions in building the autoregressive models. The classical (econometric) models described later do not receive these twelve lags directly. They instead receive the three-term design vector of the heterogeneous autoregressive (HAR) model of Corsi (2009) — the most recent month, a three-month moving average and a twelve-month moving average — which is the monthly analogue of Corsi's daily, weekly and monthly components and approximates the long-memory decay of volatility with three parameters instead of twelve. The neural network and the

BART models receive the twelve individual lags as their shrinkage prior can discover which lags matter, and Corsi's weighting is one of the hypotheses being tested rather than a maintained assumption.

The second block adds seven city-specific quantities to the autoregressive one, constructed from the same daily settlement-price returns as the realized variance itself, each capturing an aspect of the within-month return distribution. Good and bad realized volatility split the sum of squared daily returns according to the sign of the return, separating variation generated by rising prices from variation generated by falling prices, a distinction that matters since the housing literature documents an asymmetric response to shocks in recessions relative to expansions (Cheng and Chiu, 2020). Upside and downside truncated variation isolates the contribution of daily moves large enough to be treated as jumps rather than diffusive variation, so that a month whose volatility came from one large repricing is distinguished from a month of uniformly elevated turnover. A leverage term captures the covariation between the return and its own volatility, the housing analogue of the negative return-volatility relationship that motivates asymmetric volatility models generally (Nelson, 1991), while a realized skewness and realized kurtosis variables summarize the third and fourth moments of the daily return distribution within the month, providing information on the shape of the within-month distribution The aforementioned seven variables are lagged by one month, so that no quantity can contaminate the target, even the one-month-ahead forecast.

The third block adds eight principal components extracted from FRED-MD[4], the monthly macroeconomic database of McCracken and Ng (2016), which assembles roughly one hundred and thirty U.S. series covering production, employment, prices, interest rates, money and credit aggregates, and is maintained in real time so that data revisions are handled consistently. The components are national, taking a single value at each date, and are common to all ten cities. Their inclusion tests a specific proposition; a city's future volatility depends on the aggregate state of the economy over and above its own history and its own realized shocks. Eight principal components is the standard device in this literature, since they explain almost 93% of the total variation (Gupta et al., 2023).

The fourth and final block replaces the macroeconomic state with uncertainty about it, adding the six macroeconomic and financial uncertainty indices of Jurado, Ludvigson and Ng[5] (2015), each measured at forecast horizons of one, three and twelve months. These

[4] The database and its documentation are available at https://research.stlouisfed.org/econ/mccracken/fred-databases/.

[5] The indices are updated and distributed at https://www.sydneyludvigson.com/macro-and-financial-uncertainty-indexes.

indices are constructed as the conditional volatility of the purely unforecastable component of a large panel of macroeconomic and financial series and so measure how uncertain the environment is rather than what state it is in — which is precisely the distinction between this block and the previous one. Uncertainty is the more natural conditioning variable for a second-moment forecast since it has documented predictive content for housing volatility specifically, and not merely for housing returns (Strobel, Nguyen Thanh and Lee, 2020), while a rise in the one-month uncertainty that leaves the twelve-month index unchanged is a transitory disturbance, unlike a parallel shift across all three horizons that is a change in the perceived persistence of the environment. Table 2 summarizes the construction. The important structural point is that the last two blocks are parallel branches rather than successive steps, answering the question: what matters nationally is the state of the macroeconomy or the uncertainty surrounding it?

Table 2. Feature set construction

Feature set	Relation to base	Columns	Contents
AR	base set	12	Twelve monthly lags of log realized volatility
Shocks	AR + 7	19	Good and bad realized volatility, upside and downside truncated variation, leverage, realized skewness, realized kurtosis
PC	Shocks + 8	27	Eight principal components of the FRED-MD macroeconomic panel
MF	Shocks + 6	25	Jurado-Ludvigson-Ng macroeconomic and financial uncertainty at 1-, 3- and 12-month horizons

Note: PC and MF extend the Shocks base in parallel rather than in sequence. They are not nested with each other, and no specification uses both.

2.2 Methodological issues

All models are estimated on a rolling window of W=60 months and re-estimated at every forecast origin as the window advances. A rolling rather than expanding window is used because the relationship between housing volatility and its predictors is not stable over this sample: structural breaks in U.S. house price dynamics fall inside ordinary estimation windows rather than conveniently at their edges (Karoglou, Morley and Thomas, 2013), and the link between the national factor and observable macroeconomic shocks appears to have weakened after roughly 2014 (Gupta et al., 2023). An expanding window would force every forecast to be conditioned on the 2008–2010 collapse indefinitely. Sixty months is long enough to identify a twelve-lag autoregression with additional predictor blocks under shrinkage, and short enough that a regime change is fully reflected in the estimation sample within five years. Forecasts are produced at horizons of one, six and twelve months, and every model is estimated as a direct forecast rather than an iterated

one. All series are standardized strictly within the training window in force at each origin — city-level series city by city and national series by their own moments, so that no future or cross-city information enters any transformation — and every statistic in Section 3 is reported after back-transformation to the original log realized-volatility units.

The benchmark (econometric) set of our paper is constructed to span the full pooling-versus-heterogeneity spectrum, since that trade-off is what governs forecast accuracy in short panels (Pesaran, Pick and Timmermann, 2026). Pooling ten cities into a single regression buys precision at the cost of bias if the underlying coefficients differ; estimating ten separate regressions removes the bias but leaves each equation with sixty observations. Every intermediate position is represented, and all of them share the following mean equation (1):

$$y_{c,t} = \alpha_c + x_{c,t}^{\top}\varphi_c + f_{c,t}^{\top}\beta_c + \sum_{\ell=0}^{L} \bar{z}_{t-\ell}^{\top}\,\delta_\ell + \varepsilon_{c,t} \qquad (1)$$

Here $y_{c,t}$ is log realized volatility in city c at date t, and $\varepsilon_{c,t}$ is the forecast error. The regressors are $x_{c,t}$, the three-term HAR design vector holding the one-month, three-month and twelve-month moving averages of $y_{c,t}$; $f_{c,t}$, the additional predictor block in force, taken from the Shocks, PC or MF sets; and $\bar{z}_{t-\ell}$, the cross-sectional average across the ten cities of the dependent variable and the regressors at lag ℓ, included to capture the common national component. The coefficients they carry are α_c, a city-specific intercept; φ_c, the vector of HAR loadings on own past volatility; β_c, the loadings on the additional predictor block; and δ_ℓ, the loadings on the cross-sectional averages at each lag $\ell = 0, \ldots, L$. The individual estimators are restrictions on these coefficients, obtained by varying which are allowed to differ across cities and which are held common.

The **pooled Heterogeneous autoregressive model** (HAR; Corsi, 2009) model imposes $\alpha_c = \alpha$, $\varphi_c = \varphi$ and $\beta_c = \beta$ for all cities and $L = 0$ so that the cross-sectional average term is excluded, using the autoregressive set alone. It is the most heavily restricted member of the set, and for that reason it serves as the baseline against which every other model's density accuracy is reported. Its feature-augmented variants add the Shocks, PC and MF blocks in turn under the same pooling restriction and isolate the value of the additional information from the value of relaxing homogeneity. A **panel HAR-GARCH** specification pairs the mean-group mean equation with a city-specific generalised autoregressive conditional heteroskedasticity model of order (1,1) for the error variance. This is the one classical benchmark whose conditional variance moves over time, and it is included so that the gains attributed later to the score-driven variance equation cannot be dismissed as gains from having any time-varying variance at all.

Between these poles sit three estimators that relax homogeneity in different ways. The **mean-group** (MG) estimator (Pesaran and Smith, 1995) fits equation (1) city by city and averages the coefficients; it guards against the inconsistency that pooling a genuinely

heterogeneous dynamic panel induces through the lagged dependent variable. The **pooled mean-group** (PMG) estimator (Pesaran, Shin and Smith, 1999) takes the middle ground, holding the predictor-block coefficients $\beta_c = \beta$ common while letting the intercept, autoregressive dynamics and residual variance differ — the housing structure in which markets share a long-run response to national conditions but keep their own short-run dynamics. The **common correlated effects** estimator (CCE; Pesaran, 2006) instead targets the unobserved national factor that makes pooled residuals cross-sectionally correlated, augmenting the regression with cross-sectional averages that span the factor space; its **dynamic extension** (DCCE; Chudik and Pesaran, 2015) adds three lags of those averages. CCE and DCCE are the classical instrument for the same national component the PC block represents explicitly, and the contrast is informative — one absorbs the factor without estimating it, the other estimates it and lets the model choose whether to use it.

As a nonlinear, non-panel benchmark we include **Bayesian additive regression trees** (BART; Chipman, George and McCulloch, 2010), which model the conditional mean as a sum of shallow regularized trees. BART is the right control for the BNN because it shares a flexible mean disciplined by a prior rather than a validation set, but differs in exactly the two respects this paper argues are decisive: it borrows no strength across the panel beyond a city identifier, and its predictive variance is a single homoscedastic scale per window, with no counterpart to the city-specific score-driven variance developed below. Given the identical feature matrix, standardization and direct-forecast targets, the comparison isolates these two differences rather than the information set.

All mean equations are estimated either under the quasi-likelihood loss (QLIKE) or least squares. The distinction is specific to volatility forecasting and motivates one of the two training regimes carried throughout the paper. Realized volatility is not the object of interest but an imperfect estimate of it, and Patton (2011) shows that against such a proxy most loss functions rank forecasts by the noise in the proxy rather than their accuracy; only a restricted class — of which squared error and QLIKE are the leading members — leaves the ranking unchanged in expectation. Between the two, QLIKE penalizes proportional rather than squared errors, so under-predicting a quiet month by a factor of two is penalized as heavily as the same factor in a turbulent one, whereas squared-error loss is dominated by crisis months — the behavior that makes a volatility model fit the 2008–2010 collapse well and everything else poorly. Writing $u_{c,t} = y_{c,t} - \mu_{c,t}$ for the residual in logarithms, the loss and its estimating equation are

$$QLIKE(u) = e^{u} - u - 1 \tag{2}$$

$$\varphi^{(k+1)} = \left(X^{\top} W^{(k)} X\right)^{-1} X^{\top} W^{(k)} z^{(k)}, \tag{3}$$

Where $w_t^{(k)} = e^{u_t^{(k)}}$ and $z_t^{(k)} = \mu_t^{(k)} + 1 - e^{-u_t^{(k)}}$

Equation (2) is asymmetric — growing exponentially for over-prediction but only linearly for under-prediction, the correct shape for a quantity bounded below and right-skewed — and equation (3) is the iteratively reweighted least-squares scheme that minimizes it, progressively down-weighting the months where the model most over-predicts rather than letting them dictate the fit. Every specification, classical and Bayesian, is estimated twice under regimes that differ only in this loss, never in the predictors or structure. The quasi-likelihood regime fits the mean by equation (3) and draws the variance parameters under a QLIKE-weighted Metropolis–Hastings step; the least-squares regime replaces these with an ordinary-least-squares update, an exact multivariate-normal draw and an unweighted Student-t variance recursion, with the panel HAR-GARCH benchmark re-estimated by Gaussian maximum likelihood. The two are not nested and neither is a robustness check on the other: they optimize different objectives — proportional accuracy across all conditions versus likelihood under an assumed error distribution — and Section 3 shows they win on different criteria, which is why both are carried through the whole exercise.

The aforementioned classical benchmarks share three restrictions. They are linear in the predictors, their conditional variance is either constant or, in the single HAR-GARCH case, driven by squared residuals alone and the choice of which predictors to include is made by the researcher in advance, one specification at a time. The model developed here relaxes all three. Its point of departure is the single-series Bayesian neural network of Hauzenberger et al. (2025), which was designed for exactly the data configuration encountered here — few observations, many potentially relevant series, and genuine temporal dependence — and which departs from the deep-learning tradition in that the network is deliberately small and is disciplined by a prior rather than by a validation set the sample cannot spare. Three components of that template are replaced: the shrinkage prior, the treatment of the activation function, and the error variance process. We first fix notation for the neural network. For city c at time t, let $x_{c,t} \in \mathbb{R}^K$ be the feature vector of the set in force. A network with L hidden layers of J nodes maps $x_{c,t}$ to a conditional mean through $h_{c,t}^{(1)} = \phi_t\big(W_1 x_{c,t} + b_1\big), \qquad h_{c,t}^{(\ell)} = \phi_t\Big(W_\ell h_{c,t}^{(\ell-1)} + b_\ell\Big), \quad \ell = 2, \dots, L$

under $\mu_{c,t} = \beta_0 + \beta^\top h_{c,t}^{(L)} + \rho^\top x_{c,t}, y_{c,t} = \mu_{c,t} + \varepsilon_{c,t}$ with layer weights W_ℓ and biases b_ℓ, output weights β, a skip-layer linear term ρ, activation ϕ_t — below a time-varying mixture rather than a fixed function — and an error $\varepsilon_{c,t}$ following a score-driven variance process. The three components that depart from the single-series template — the shrinkage prior on W_ℓ, β, ρ, the activation ϕ_t, and the distribution of $\varepsilon_{c,t}$ — are described in turn.

A network of this size carries several hundred weights against a sixty-month window, so without regularization it interpolates the training data and forecasts badly, while cross-

validation on short, serially dependent samples and an explicit specification search over predictor blocks are both unavailable. The standard Bayesian remedy, and the one adopted here is a global–local shrinkage prior: a single global scale controls overall sparsity while one local scale per weight lets genuinely informative coefficients escape it, giving a continuous analogue of variable selection with no discrete choice and no tuning constants (Carvalho, Polson and Scott, 2010). We use the regularized, or Finnish, variant of Piironen and Vehtari (2017), whose slab component places a floor under how far the largest coefficients can be shrunk and, through a neuron-level scale, allows the prior to switch off an entire hidden node — so that the network's effective width is set by the data rather than by the architecture grid, which is why forecast accuracy turns out in Section 3 to be almost invariant to the nominal number of nodes and layers. The shrinkage machinery is standard rather than a contribution of this paper; the exact prior, its hyperpriors and its slab calibration are given in Appendix B.1.

The activation function is not fixed in advance but treated as a time-varying mixture over four candidate shapes (hyperbolic tangent, rectified linear, logistic and leaky rectified linear), with mixture weights driven by a score recursion of their own so that the curvature of the predictor–volatility relationship can adapt between calm and crisis periods. In the event this device does not differentiate the results — the mixture entropy sits at its maximum throughout the sample (Section 3.6) — so we retain it as a robustness feature rather than a contribution; its full specification and recursion are in Appendix B.2. The alternative used here is a score-driven, city-specific exponential GARCH in the Beta-t form of Harvey and Chakravarty (2008), belonging to the generalized autoregressive score family of Creal, Koopman and Lucas (2013) and Harvey (2013) and used for time-varying risk prices by Umlandt (2023)[6]. Writing $\nu_{c,t}$ for the log-variance state of city c and $\varepsilon_{c,t}$ for the model residual,

$$\nu_{c,t} = \omega_c + \beta_c\,\nu_{c,t-1} + \alpha_c\,u^*_{c,t-1} \tag{4}$$

$$\text{under } u^*_{c,t} = \frac{(d_c+1)\,z^2_{c,t}}{d_c+z^2_{c,t}} - 1, \qquad z_{c,t} = \varepsilon_{c,t}\,e^{-\nu_{c,t}/2}$$

[6] The mean equation described so far produces a point forecast. What a risk manager requires is a predictive density, and its dispersion cannot be treated as constant across a sample containing the 2008 collapse, the 2020 shutdown and the 2022 rate shock. The single-series model of Hauzenberger et al. (2025) handles this with a stochastic-volatility process — a latent autoregression driven by its own innovation. The specific choice adds one latent state per period per city to a sampler that is already expensive, and, more importantly, its updating is not tied to the model's own forecast errors: the latent variance can drift away from what the residuals imply and be corrected only slowly.

The variance state is therefore updated by the scaled score of the predictive log-likelihood, which is to say it moves in the direction that would have improved the density forecast of the observation just seen. $u^*_{c,t}$ in equation (4) is a rescaled Beta-distributed quantity and is bounded, so an extreme residual contributes a score that saturates. In a Gaussian exponential GARCH the same residual enters squared and can inflate the estimated scale for many periods afterwards — precisely the failure mode of a volatility model applied to a series with year 2008 in it. Moreover, our approach is not a convenient heuristic. Beutner, Lin and Lucas (2026) establish that the filtered path of a score-driven model converges to the Kullback–Leibler-optimal time-varying parameter path even when the model is badly misspecified, which is the relevant guarantee for a variance equation attached to a neural network whose mean specification is certainly not the true one.

The variance parameters are city-specific by design, so that a market which is quiet for years and then moves sharply need not share a persistence parameter with one that moves continuously, and the degrees-of-freedom parameter d_c is estimated rather than fixed, so each city's tail thickness is set by its own data. The priors, the informative Beta prior that concentrates persistence near 0.9, and the Metropolis–Hastings updating are given in Appendix B.3. Residuals entering both the quasi-likelihood objective and the variance recursion are weighted by an exponential decay in age, normalized to average one across the window so that the total quantity of information used is unchanged and only its distribution across time is altered:

$$w_t = \frac{e^{-\gamma(T-1-t)}}{\frac{1}{T}\sum_{\tau=0}^{T-1} e^{-\gamma(T-1-\tau)}} \tag{5}$$

The decay rate γ is not fixed a priori, since the appropriate degree of discounting depends on how unstable the recent past has been. It is treated as a parameter, shared across cities because that instability is a national phenomenon, given an exponential prior with rate 50 — a prior mean of 0.02, or very mild discounting — and updated by twenty Metropolis–Hastings steps per Gibbs sweep. It is capped at 0.2, where the earliest month in the window receives roughly one-twentieth the weight of the most recent; without the cap the sampler can drive γup until the effective sample collapses onto the final few months and the variance parameters become unidentified.

The posterior is drawn by a Gibbs sampler in which the tractable blocks are sampled directly and the hidden-layer weights by Hamiltonian Monte Carlo, which follows the geometry of the posterior rather than diffusing through it and so remains viable in the several hundred dimensions a neural network occupies (Duane et al., 1987; Neal, 2011). We use the no-U-turn variant, which sets trajectory length automatically and adapts the

step size during warm-up (Hoffman and Gelman, 2014)[7]. Each rolling window is estimated with 1,000 warm-up and 1,000 retained sweeps; the full algorithm is in Appendix B.5.

A forecasting comparison is only as informative as the criteria it is settled on, and because this paper's claim concerns predictive densities rather than point forecasts, the criteria are set out before any results are reported. The first criterion under consideration is the **average log predictive likelihood** (ALPL). Differences in ALPL are in logarithmic units and are read like log-likelihood differences of the pooled HAR baseline at the same horizon and under the same training regime. A model scores well only if it is accurate *and* honest about its own accuracy: a forecast centered close to the outcome but reported with an implausibly narrow interval is penalized for the density it fails to place where the outcome fell, and a forecast with wide intervals that never commits to anything is penalized for the density it wastes elsewhere. A capital requirement, a value-at-risk limit and a stress threshold are all statements about the tail of a distribution, and a model that ranks first on point accuracy while systematically understating its own uncertainty is worse than useless for setting them, because it is confidently wrong in exactly the states where being wrong is expensive.

On the other hand, average performance can conceal compensating errors, and a model may be well calibrated in the center of the predictive distribution while being badly calibrated in the tails that matter for risk management. Two further criteria address this. The first is the **quantile score** (QS), evaluated at $\tau \in \{0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99\}$. The second is **value-at-risk (VaR) coverage**: the empirical frequency with which the outcome falls below the model's own predicted quantile at a stated level. Two further statistics guard against the possibility that a density advantage is an artefact of the likelihood rather than a genuinely better forecast. The **root mean squared error** (RMSE) ratio is a model's root mean squared forecast error divided by that of the pooled HAR baseline, so that the baseline takes the value one by construction and values below one indicate tighter point forecasts. The **continuous ranked probability score** (CRPS) closes that gap. A proper scoring rule that integrates the squared distance between the predictive and empirical cumulative distribution functions, it evaluates the whole distribution rather than its center alone, is reported in the outcome's own units, and rewards lower values (Gneiting and Raftery, 2007). We compute it in sample form from the posterior draws (Appendix B.4), averaged over cities and dates. Naturally, the valuation of each mode and the comparison between them is based on a joint examination based on all criteria.

[7] Implemented in the probabilistic programming language of Carpenter et al. (2017).

Statistical accuracy is not the sole object of the exercise. The applications above — mortgage risk pricing, hedging of mortgage-backed security portfolios and macroprudential capital setting — turn on whether a better volatility forecast changes what an institution must hold, a question answered in units of capital rather than log-likelihood. The economic evaluation therefore reports the **width of the ninety-five per cent predictive interval**, expressed as a percentage reduction relative to the pooled HAR baseline, alongside its realized coverage. The logic is that of value-at-risk-proportional capital rules, in which required capital scales with an estimated risk measure at a fixed confidence level: a narrower interval at comparable coverage delivers the same protection at lower capital cost, so the width reduction proxies the saving. Width is meaningful only jointly with **coverage**, since a model can always narrow its intervals by understating its uncertainty, and a narrow interval that under-covers is not cheaper protection but absent protection. Any specification whose realized coverage falls outside a tolerance band around the nominal level is therefore flagged, and its width advantage reported as miscalibration rather than an economic gain.

Pairwise model comparisons use the Diebold and Mariano (1995) statistic, but testing across twenty-nine panel specifications invites a multiple-comparison problem, so the comparison is examined by the model confidence set (MCS) of Hansen, Lunde and Nason (2011) in Appendix A3. Differences in realized coverage and exceedance rates between the two training objectives are assessed by a stationary block bootstrap (Politis and Romano, 1994) applied to months rather than observations, with the cross-section retained within each resampled month, using 50,000 replications and an expected block length of twelve months. Further evaluation detail is in Appendix B.4.

3. Empirical findings

3.1 Results under quasi-likelihood training

The forecasting exercise evaluates three model families: the classical panel HAR estimators — the pooled, mean-group (MG), pooled-mean-group (PMG), common-correlated-effects (CCE) and dynamic CCE (DCCE) variants, together with a panel HAR-GARCH — Bayesian additive regression trees, and the panel Bayesian neural network, each across four feature sets, three forecast horizons and two training regimes. The classical family contributes thirteen specifications and BART four; the BNN is estimated over four architectures per feature set[8]. Counting every combination that is actually fitted, the comparison rests on twenty-nine distinct panel models per horizon per regime, together with the four BART benchmarks, evaluated on identical data, standardization and direct-forecast targets so that differences reflect the model rather than

[8] 1 or 2 hidden layers and 10 or 20 neurons per hidden layer. More details in Appendix A1.

the information set. The two regimes are reported separately, because each model's density accuracy is measured against a pooled HAR baseline re-estimated under that same regime. To keep the tables legible each reports the leading architecture per feature family; the complete model-by-model results are in Appendix Tables A1.

Table 3. Density and point accuracy at the one-month horizon, quasi-likelihood training

Model	ΔALPL	RMSE	CRPS	Cov. 95%	VaR 5%	Width 95%
Pooled HAR	+0.000†	2.172***	1.237***	0.923	0.137	7.900
Panel HAR-GARCH	-0.011	2.199***	1.248***	0.946	0.101	8.910
BNN-PC	-0.021	1.927***	1.153†	0.995	0.011	13.010
Pooled HAR-Shocks	-0.036***	2.215***	1.269***	0.902	0.154	7.622
BNN-MF	-0.037	1.910†	1.168**	0.996	0.011	13.600
BNN-AR	-0.041	1.941**	1.176***	0.993	0.017	13.448
MG-HAR	-0.052***	2.199***	1.267***	0.894	0.162	7.360
Pooled HAR-PC	-0.061*	2.219***	1.279***	0.885	0.144	7.123
BNN-Shocks	-0.069**	1.977***	1.211***	0.996	0.014	14.167
Pooled HAR-MF	-0.086**	2.259***	1.308***	0.892	0.161	7.072
PMG-HAR-PC	-0.149***	2.261***	1.326***	0.842	0.163	6.721
PMG-HAR-MF	-0.162***	2.282***	1.341***	0.844	0.187	6.683
CCE-HAR	-0.204***	2.267***	1.329***	0.802	0.189	5.815
CCE-HAR-MF	-0.211***	2.265***	1.329***	0.798	0.188	5.753
BART-PC	-0.223	2.111***	1.223***	0.803	0.253	5.337
DCCE-HAR	-0.242***	2.311***	1.357***	0.793	0.194	5.779
DCCE-HAR-MF	-0.243***	2.298***	1.350***	0.789	0.190	5.712
BART-MF	-0.286	2.178***	1.269***	0.777	0.270	5.363
BART-AR	-0.374	2.281***	1.332***	0.763	0.294	5.521
MG-HAR-MF	-0.394***	2.551***	1.520***	0.778	0.227	6.214
BART-Shocks	-0.401	2.303***	1.347***	0.758	0.291	5.522

Note: † leader by metric; asterisks give the Diebold–Mariano significance of the leader's advantage. Cov. 95% and VaR 5% are realized frequencies against nominal levels; Width 95% is the mean predictive-interval width. Leading architecture per family; full results in Appendix Table A1. Authors' calculations.

Reading down the table by family, no model separates itself from the pooled HAR baseline on density at this horizon: the entire field lies within roughly two-tenths of a log point of zero, the pooled HAR is the nominal leader, the panel HAR-GARCH is indistinguishable from it at −0.011, and the three neural families cluster immediately below, from BNN-PC at −0.021 through BNN-MF and BNN-AR to BNN-Shocks at −0.069, none separable from the leader by the Diebold–Mariano test. Read on the log predictive likelihood alone the table therefore looks like a dead heat. It is not. The point

and full-distribution columns break the tie in the BNN's favor: the neural families post the lowest root mean squared errors in the table, near 1.92 against the baseline's 2.17, and the lowest continuous ranked probability scores, near 1.15 against 1.24 — and they do so ahead of every classical and tree specification, so the advantage registers on two independent criteria that the density near-tie conceals. The BART family sits far lower on density, between −0.22 and −0.40, with coverage near 0.78 and narrow intervals near 5.4: a homoscedastic tree ensemble reweighted for the quasi-likelihood objective is poorly served by it. Beneath the neural cluster the classical panel estimators deteriorate monotonically on every column, from the pooled and mean-group specifications through the common-correlated-effects variants to the dynamic CCE estimators at −0.24, whose coverage falls to 0.79 and whose exceedance rate rises above 0.18. These pairwise Diebold–Mariano comparisons have a joint counterpart in the model confidence set of Appendix A3, which confirms that the leaders are statistically indistinguishable from most of the field. Thus, pooled HAR leads on density while the BNN leads on RMSE and CRPS.

The calibration columns then say what kind of advantage it is. The neural densities are the most conservative in the table — central coverage near 99.5 per cent, five per cent value-at-risk exceedance near one per cent, and the widest intervals, near 13 to 14 log units — which is the behavior the quasi-likelihood loss was selected to produce and the opposite of the classical estimators, whose 79 to 92 per cent coverage and 14 to 19 per cent exceedance understate risk. So, at one month the reading is not that the conditional density is unpredictable — the point and distribution scores show the BNN captures it best — but that the model's edge is bought as conservatism: it insures against the tail by widening, holding coverage well above nominal. Conservatism of this degree is itself a calibration failure. A ninety-five per cent interval that contains 99.5 per cent of outcomes misses its nominal level by four and a half points, and it does so by carrying intervals of thirteen to fourteen log units against the pooled HAR's 7.9, so the protection is bought with roughly seventy per cent more width at the same stated confidence. The distinction that matters for the applications of Section 1 is not that one model is calibrated and the others are not, since none of them is, but that the errors point in opposite directions and only one direction is penalised. A supervisory backtest counts breaches of the stated quantile and is silent on unused width, so a model that over-covers holds more capital than it needs while a model that under-covers attracts a multiplier. The classical estimators' 79 to 92 per cent coverage and 14 to 19 per cent exceedance rates fall on the penalised side of that asymmetry.

Table 4. Density and point accuracy at the six-month horizon, quasi-likelihood training

Model	ΔALPL	RMSE	CRPS	Cov. 95%	VaR 5%	Width 95%
BNN-MF	+0.015†	2.181	1.286**	0.984	0.025	13.854

Model	ΔALPL	RMSE	CRPS	Cov. 95%	VaR 5%	Width 95%
BNN-AR	+0.009	2.079†	1.275†	0.992	0.016	14.793
BNN-Shocks	+0.006	2.092	1.285*	0.993	0.016	14.964
Pooled HAR	+0.000	2.364	1.376***	0.898	0.166	7.896
BNN-PC	-0.003	2.290**	1.326	0.982	0.035	13.881
Pooled HAR-Shocks	-0.006	2.330***	1.363***	0.886	0.188	7.472
Panel HAR-GARCH	-0.028	2.433***	1.407***	0.912	0.134	8.910
CCE-HAR	-0.050	2.431***	1.422***	0.882	0.180	7.724
MG-HAR	-0.104*	2.433***	1.439***	0.854	0.195	7.236
DCCE-HAR	-0.137**	2.545***	1.498***	0.850	0.210	7.501
BART-PC	-0.264	2.263***	1.330***	0.763	0.272	5.368
BART-MF	-0.350	2.341***	1.384***	0.746	0.303	5.347
PMG-HAR-MF	-0.391***	6.365***	2.095***	0.805	0.233	9.293
CCE-HAR-MF	-0.399***	2.933***	1.715***	0.785	0.237	6.717
DCCE-HAR-MF	-0.401***	2.761***	1.683***	0.765	0.227	6.506
BART-Shocks	-0.456	2.465***	1.478***	0.732	0.318	5.572
BART-AR	-0.470	2.476***	1.487***	0.729	0.326	5.554
Pooled HAR-MF	-0.507**	3.333***	1.859***	0.784	0.240	6.849
Pooled HAR-PC	-0.721***	3.468***	2.056***	0.732	0.265	6.874
PMG-HAR-PC	-0.743***	4.564***	2.315***	0.729	0.333	7.776
MG-HAR-MF	-1.404***	4.779***	2.587***	0.616	0.305	5.852

Note: as Table 3. Authors' calculations.

At six months (Table 4) the ranking by average log predictive likelihood places the neural families first, led by BNN-MF at +0.015, with BNN-AR and BNN-Shocks at +0.009 and +0.006. The Diebold–Mariano test separates the leader from roughly a third of the field, so these density differences are not individually significant against most competitors and the ordering at the top should be read as a near-tie. The point and full-distribution scores are more separated, with the neural families recording the lowest root mean squared errors (near 2.08, versus 2.36 for the pooled HAR) and the lowest CRPS (near 1.28, versus 1.38), ahead of every classical and BART specification. On calibration the neural densities cover 0.98 to 0.99 against a nominal 0.95, with five per cent value-at-risk exceedances of 0.02 to 0.04 and interval widths of 14 to 15 log units — over-coverage achieved by widening. The classical estimators rank below on every column: the pooled HAR covers 0.90 with an exceedance rate of 0.17, and the CCE, MG and DCCE families range from −0.05 to −0.14 with coverage in the mid-0.80s. BART is the lowest, −0.26 to −0.47, covering 0.73.

Thus, the density gains over a pooled autoregression are small and, by the DM test, largely unresolved, so no strong ranking claim is warranted at this horizon on density alone. The separation that does hold is in point accuracy, full-distribution score and

calibration: the neural specifications attain lower RMSE and CRPS and coverage within tolerance of nominal, whereas the classical estimators under-cover by 5 to 10 percentage points and would, on their exceedance counts, attract a back-testing penalty.

Table 5. Density and point accuracy at the twelve-month horizon, quasi-likelihood training

Model	ΔALPL	RMSE	CRPS	Cov. 95%	VaR 5%	Width 95%
BNN-PC	+0.063†	2.226*	1.310	0.987	0.016	13.690
BNN-Shocks	+0.048	2.129	1.302	0.988	0.018	14.991
BNN-AR	+0.047	2.116†	1.297†	0.988	0.018	14.990
Pooled HAR	+0.000	2.418***	1.420***	0.873	0.174	7.431
CCE-HAR	-0.012	2.426***	1.428***	0.873	0.186	7.347
Pooled HAR-Shocks	-0.028	2.426***	1.433***	0.854	0.191	7.099
BNN-MF	-0.051	2.707***	1.475	0.969	0.037	14.796
Panel HAR-GARCH	-0.068**	2.583***	1.500***	0.875	0.151	8.404
Pooled HAR-MF	-0.142*	2.571***	1.519***	0.806	0.199	6.803
DCCE-HAR	-0.147***	2.583***	1.548***	0.827	0.199	7.007
MG-HAR	-0.202***	2.583***	1.544***	0.795	0.223	6.685
CCE-HAR-MF	-0.227**	2.676***	1.597***	0.782	0.234	6.751
BART-PC	-0.249	2.261***	1.332	0.756	0.280	5.262
BART-MF	-0.349	2.354***	1.404**	0.737	0.304	5.288
BART-Shocks	-0.475	2.487***	1.504***	0.716	0.318	5.454
BART-AR	-0.515	2.525***	1.530***	0.710	0.325	5.479
PMG-HAR-MF	-0.887***	4.205***	3.698***	0.746	0.278	7.693
DCCE-HAR-MF	-0.920**	4.587***	2.354***	0.704	0.271	6.455
Pooled HAR-PC	-1.208***	4.513***	2.561***	0.642	0.350	6.387
PMG-HAR-PC	-1.424***	9.434***	4.068***	0.640	0.372	17.516
MG-HAR-MF	-2.171***	6.379***	3.283***	0.519	0.280	5.492

Note: as Table 3. Authors' calculations.

At twelve months reported in Table 5, the density advantage is at its largest, though it remains modest in absolute terms, and its source shifts. The principal-component BNN now leads at +0.063, ahead of the autoregressive and shock BNN at +0.047 and +0.048 — the macroeconomic-state block overtaking the within-month shock measures that led at no shorter horizon, consistent with the macroeconomic state being persistent while realized shocks decay within a quarter (Gupta et al., 2023; Tidwell et al., 2023). As before, the ranking read off the log predictive likelihood alone understates the separation and the models at the top are not statistically different between them. On the RMSE loss, the leading BNN-AR is more accurate than every classical estimator at the one per cent level and than every BART variant, with the shock and PC BNNs statistically indistinguishable from it. On CRPS the same holds against all four classical estimators,

with the single exception that BART-PC (1.332) is not separable from the leader (1.297), so on the full-distribution score one tree specification matches the BNN at this horizon. The point-accuracy gap over the pooled HAR is roughly a third of a log-RV unit in RMSE and comparable in CRPS, and it is these two criteria, not the density score, that the data resolve most cleanly.

The calibration columns show the advantage to be conservative, and the gap between the neural and classical families is wider here than at any shorter horizon. The neural densities cover 0.987 to 0.988 against a nominal 0.95, with five per cent value-at-risk exceedances of 0.016 to 0.018 and interval widths near 14 to 15 log units — over-coverage bought by widening. The classical estimators move the opposite way as the horizon lengthens: the pooled HAR and the static common-correlated-effects estimator hold near-nominal density at zero and −0.012 but cover only 0.873, while the mean-group and dynamic CCE families fall to −0.202 and −0.147 with coverage of 0.795 and 0.827 and exceedance rates above 0.20 — four times nominal. At the ninety-nine per cent level that record would place their internal models in the Basel red zone and attract the maximum capital add-on, where the neural specifications remain green. BART again posts the narrowest intervals, near 5.3, at coverage of 0.71 to 0.76 with narrowness that reflects under-coverage rather than efficiency. The reading at twelve months is therefore that the level of volatility is barely more predictable than a pooled autoregression makes it — the density gains are small — but that the BNNs attain significantly lower point and full-distribution loss and hold supervisory-grade coverage precisely where the classical estimators fail it, and that failure is at its most pronounced at the longest horizon.

3.2 Results under least-squares training

For comparison reasons with the existing literature, we repeat the forecasting exercise using least-squares training. In Table 6 we report the evaluation metrics at the one-month horizon.

Table 6. Density and point accuracy at the one-month horizon, least-squares training

Model	ΔALPL	RMSE	CRPS	Cov. 95%	VaR 5%	Width 95%
BART-AR	+0.028†	1.871	1.065†	0.919	0.106	6.596
BNN-MF	+0.023	1.848†	1.065†	0.945	0.075	7.215
BNN-AR	+0.020	1.876**	1.072	0.948	0.080	7.531
BART-Shocks	+0.019**	1.885***	1.074***	0.913	0.106	6.590
BART-PC	+0.018	1.871*	1.070	0.911	0.091	6.372
BART-MF	+0.014	1.879**	1.078**	0.910	0.092	6.440
BNN-Shocks	+0.005	1.898**	1.086***	0.941	0.088	7.464
Pooled HAR	+0.000	1.939***	1.101***	0.968	0.064	8.373
BNN-PC	-0.005	1.891***	1.091**	0.930	0.074	7.117
Pooled HAR-Shocks	-0.018	1.977***	1.123***	0.959	0.070	8.238

Model	ΔALPL	RMSE	CRPS	Cov. 95%	VaR 5%	Width 95%
MG-HAR	-0.038	2.008***	1.144***	0.954	0.073	8.126
Panel HAR-GARCH	-0.040	2.008***	1.144***	0.948	0.073	8.111
Pooled HAR-MF	-0.071	2.075***	1.183***	0.946	0.071	7.990
Pooled HAR-PC	-0.102	2.155***	1.217***	0.933	0.064	7.953
PMG-HAR-MF	-0.103	2.112***	1.211***	0.925	0.082	7.740
PMG-HAR-PC	-0.167	2.245***	1.280***	0.909	0.066	7.697
CCE-HAR	-0.296	2.268***	1.331***	0.816	0.146	6.171
CCE-HAR-MF	-0.303	2.280***	1.336***	0.819	0.146	6.160
MG-HAR-MF	-0.313	2.413***	1.407***	0.854	0.125	7.192
DCCE-HAR-MF	-0.347	2.332***	1.371***	0.808	0.150	6.128
DCCE-HAR	-0.348	2.335***	1.373***	0.808	0.150	6.126

Note: as Table 3. Authors' calculations.

Under least-squares training the one-month ordering tightens to near-indistinguishability at the top. BART-AR is the nominal leader at +0.028, the neural families follow at +0.023 (BNN-MF) and +0.020 (BNN-AR), and the remaining BART variants sit between them, the whole leading group spanning six thousandths of a log point with none separable from the leader by the Diebold–Mariano test. The point and full-distribution columns tell the same story at the top and a sharper one below it. Against the classical estimators the leading BNN is significantly more accurate on both RMSE and CRPS at the one per cent level — 1.848 and 1.065, against the pooled HAR's 1.939 and 1.101 — but against BART it is not, since BART-AR is statistically level with the best BNN on RMSE and is itself the nominal leader on CRPS. So, under this objective the BNN's edge is over the panel-HAR family, not over the nonlinear tree competitor, which matches it on every accuracy criterion. What distinguishes the two training regimes is therefore not accuracy but calibration.

The calibration columns sharpen the point and correct a natural misreading. The pooled HAR posts the highest central coverage in the table, 0.968, but that is over-coverage, not superior calibration: nominal is 0.95, the leading BNN sits almost exactly on it at 0.945, and the HAR reaches its higher number only by widening its intervals to 8.37 log units against the BNN's 7.21 — more than a sixth wider, which at a fixed confidence level is more capital held for the same nominal protection. Where the BNN is genuinely less conservative is the extreme tail: its five per cent value-at-risk exceedance runs at 0.075 against the HAR's 0.064, so it delivers a declared one-in-twenty loss closer to one in thirteen. That single number is the economically material line of the table. It is also the mirror image of Table 3, where the very same BNNs on the very same data produced over-conservative tails under the quasi-likelihood objective. This contrast — identical models, opposite direction of miscalibration, set entirely by the estimation loss — is the paper's central finding on training objectives, and it is not one the literature anticipates:

the choice between a proxy-robust loss and a likelihood is normally cast as a question of estimator consistency (Patton, 2011), whereas here it determines which of two economically opposite errors the user inherits, and Section 3.6 shows the difference is large enough to move a model between Basel capital zones.

Table 7. Density and point accuracy at the six-month horizon, least-squares training

Model	ΔALPL	RMSE	CRPS	Cov. 95%	VaR 5%	Width 95%
BART-MF	+0.092†	1.928†	1.107†	0.894	0.108	6.353
BART-PC	+0.070	1.946	1.117	0.892	0.109	6.343
BNN-Shocks	+0.023	2.060***	1.190***	0.937	0.094	7.654
BNN-AR	+0.021	2.065***	1.192***	0.935	0.088	7.805
BART-Shocks	+0.019**	2.048***	1.181***	0.900	0.120	6.659
BART-AR	+0.014**	2.059***	1.188***	0.893	0.113	6.706
Pooled HAR	+0.000	2.148***	1.234***	0.956	0.080	8.487
Pooled HAR-Shocks	-0.003	2.145***	1.235***	0.948	0.082	8.115
BNN-MF	-0.019	2.119**	1.204**	0.918	0.120	7.174
CCE-HAR	-0.020	2.188***	1.257***	0.949	0.087	8.425
DCCE-HAR	-0.071	2.281***	1.315***	0.927	0.104	8.288
MG-HAR	-0.072	2.259***	1.303***	0.924	0.093	8.163
Panel HAR-GARCH	-0.091	2.259***	1.309***	0.914	0.100	8.120
PMG-HAR-MF	-0.155	2.449***	1.409***	0.919	0.103	8.191
CCE-HAR-MF	-0.369	2.936***	1.633***	0.860	0.141	7.752
DCCE-HAR-MF	-0.376	2.755***	1.617***	0.816	0.149	7.543
BNN-PC	-0.392	2.772***	1.549***	0.839	0.156	7.007
PMG-HAR-PC	-0.440	3.027***	1.768***	0.850	0.146	8.620
Pooled HAR-MF	-0.442	3.257***	1.741***	0.861	0.144	7.858
Pooled HAR-PC	-0.572	3.117***	1.829***	0.803	0.167	7.766
MG-HAR-MF	-1.168	4.810***	2.466***	0.699	0.212	7.030

Note: as Table 3. Authors' calculations.

At six months under least-squares training (Table 7) the leading models on accuracy are homoscedastic tree ensembles. BART-MF and BART-PC hold the lowest root mean squared errors in the table, 1.928 and 1.946 against the pooled HAR's 2.148, and the lowest CRPS, 1.107 and 1.117; the Diebold–Mariano test does not separate the two from each other, but places every neural family significantly behind — BNN-Shocks and BNN-AR at the one per cent level near 2.06, the principal-component BNN collapsing to 2.77 and −0.392 in ΔALPL. On point and full-distribution accuracy, then, BART leads and the BNN's trail, clearly for BNN-PC and significantly for the rest. Calibration, however, runs the other way. BART attains its accuracy at coverage of 0.89, six points below nominal, with a five per cent value-at-risk exceedance of 0.11 — a declared one-in-twenty loss arriving closer to one in nine — whereas the neural densities cover 0.93 to

0.94 with exceedances near 0.09, closer to nominal from below, and the pooled HAR is closest of all at 0.956. So, the tree ensemble's accuracy lead is bought with a hotter tail, not delivered alongside good calibration. In this longer horizon under a least-squares objective where the conditional variance is close to constant, BART forecasts the level more accurately than a BNN, but it remains an asset for calibration, and the gap is large. The BNN's intervals cover 0.937 against BART's 0.894, cutting the shortfall from nominal by more than three-quarters (a 1.3-point gap against BART's 5.6), and its five per cent exceedance runs at 0.094 against BART's 0.108, so that in a 250-observation supervisory window BART breaches about 27 times. The same fixed-variance simplicity that wins BART the accuracy race is what leaves its tail materially hotter, so the variance machinery earns its place precisely where it costs accuracy.

Table 8. Density and point accuracy at the twelve-month horizon, least-squares training

Model	ΔALPL	RMSE	CRPS	Cov. 95%	VaR 5%	Width 95%
BART-MF	+0.109†	1.958	1.119†	0.875	0.127	6.251
BART-PC	+0.107	1.956†	1.122	0.878	0.101	6.231
Pooled HAR-MF	+0.006	2.208***	1.268***	0.904	0.080	7.624
CCE-HAR	+0.003	2.246***	1.288***	0.920	0.077	8.136
BART-Shocks	+0.002*	2.136***	1.232***	0.860	0.124	6.493
Pooled HAR	+0.000	2.253***	1.292***	0.921	0.075	8.140
Pooled HAR-Shocks	-0.004	2.257***	1.293***	0.918	0.090	7.908
BART-AR	-0.023*	2.179***	1.260***	0.860	0.118	6.617
BNN-Shocks	-0.025	2.211***	1.284***	0.899	0.114	7.443
CCE-HAR-MF	-0.030	2.271***	1.308***	0.892	0.102	7.598
BNN-AR	-0.030	2.228***	1.296***	0.898	0.107	7.663
MG-HAR	-0.134	2.439***	1.412***	0.872	0.110	7.774
DCCE-HAR	-0.155	2.482***	1.455***	0.871	0.114	7.689
Panel HAR-GARCH	-0.158	2.439***	1.415***	0.865	0.115	7.776
BNN-MF	-0.228	2.904**	1.471**	0.869	0.144	6.931
PMG-HAR-MF	-0.382	7.134***	2.364***	0.857	0.152	10.462
BNN-PC	-0.530	3.006***	1.722***	0.785	0.174	6.877
PMG-HAR-PC	-0.739	7.319***	2.878***	0.818	0.220	13.743
DCCE-HAR-MF	-0.744	3.996***	2.079***	0.778	0.193	7.138
Pooled HAR-PC	-1.163	4.457***	2.586***	0.702	0.272	7.546
MG-HAR-MF	-2.007	5.920***	3.259***	0.568	0.218	6.544

Note: as Table 3. Authors' calculations.

At twelve months under least-squares training (Table 8) the tree ensemble's accuracy advantage widens. BART-MF and BART-PC lead at +0.109 and +0.107, with the lowest root mean squared errors in the table, 1.958 and 1.956 against the pooled HAR's 2.253,

and the Diebold–Mariano test places every neural family significantly behind — the best of them, BNN-AR and BNN-Shocks, at 2.21 to 2.23, the principal-component and macro-factor **BNNs** failing outright at 3.01 and 2.90 and ΔALPL of −0.53 and −0.23. On point and density accuracy this is the region of the design where the **BNNs** are most clearly beaten, and it is reported as such. However, calibration does not follow the same ordering, as BART buys its accuracy at coverage of 0.875, nearly seven points below nominal, with a five per cent exceedance of 0.127 — some 32 breaches in a 250-observation window against a nominal 12.5 — while the neural families cover 0.898 to 0.899 and the pooled HAR, closest of all, 0.921. So even here, where a disciplined tree prior forecasts the level best, its tail is the least well calibrated of the serious models, and the ranking on accuracy is the reverse of the ranking on coverage. Among the classical estimators the pooled HAR with the uncertainty indices and the static CCE estimator sit just above zero, at +0.006 and +0.003, while the neural families fall below the baseline on density, the best at −0.025. The next section takes up the interpretation, locating the **BNN**'s contribution at the short horizon where the second moment is predictable and explaining its retreat at the long one where a fixed-variance tree forecasts the level more accurately. Even there the **BNN** holds an edge, its variance equation delivering the better-behaved tail that BART's accuracy is bought against.

3.3 Contribution of the variance equation

The comparison with BART is a decomposition rather than a contest. Its purpose is to attribute the calibration behavior documented above to one of the two components that distinguish our specification from a conventional nonlinear forecaster, and BART is suited to it because it carries a flexible nonlinear conditional mean while holding a single homoscedastic scale per estimation window. Varying the variance process while holding the mean flexible therefore identifies what the score-driven filter contributes, and the sign of that contribution is not settled in advance: a time-varying variance improves calibration only if the second moment is genuinely predictable at the relevant horizon, and Sections 3.1 and 3.2 have already established how little of it is. Heteroskedastic extensions of tree ensembles exist, through multiplicative variance trees or a stochastic-volatility component, and a comparison against one of those would ask a different and narrower question, namely whether a score-driven filter outperforms an alternative time-varying variance specification. That question is worth asking and we do not answer it. What the present comparison establishes is that the variance specification and not the conditional mean is where the calibration difference resides.

Tables 3 to 8 show that the two models exchange the lead in root mean squared error and CRPS within a narrow band at every horizon. The BNN’s score-driven variance therefore buys no advantage in point or density accuracy over a nonlinear mean carrying a fixed per-window scale. However, the contribution appears in the tail. Figure 1 reports, city by city and horizon by horizon, the number of the one per cent value-at-risk breaches each

model commits in a 250-observation supervisory window, against the Basel green, yellow and red zones. The pattern is uniform and more pronounced under quasi-likelihood training, where the BNN sits in the green zone almost everywhere, breaching its stated tail only a handful of times per window, while BART is deep in the red, breaching many times more often than it declares. The same accuracy that the tables show to be a near-tie is here accompanied by a calibration gap that a supervisor, who reads a model through its exceedance count, would find decisive, however well it forecasts the center of the distribution.

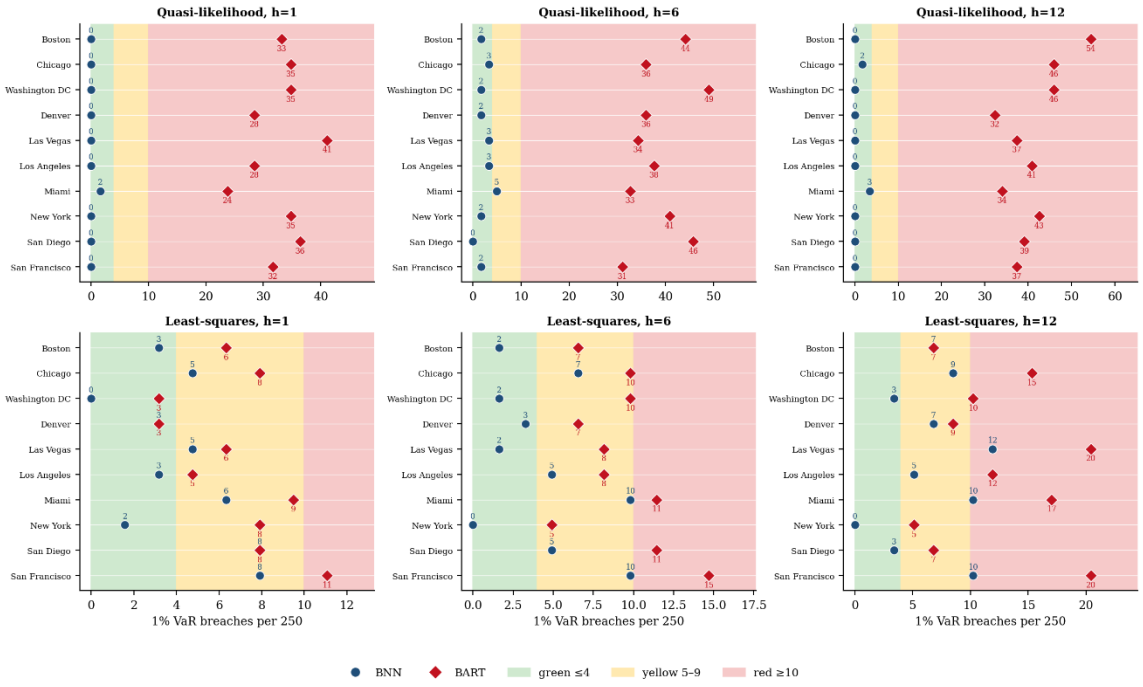


Figure 1. One per cent value-at-risk breaches per 250 observations, BNN versus BART, by city, horizon and training regime. *Each panel places the ten cities on the vertical axis and the number of breaches of the one per cent tail per 250 observations on the horizontal, with the Basel green (four or fewer), yellow (five to nine) and red (ten or more) zones shaded. Circles are the BNN, diamonds BART. A correctly calibrated one per cent tail produces 2.5 breaches per 250 observations, so points far to the right under-state risk and points at the extreme left over-state it; only the latter escapes a supervisory penalty. Authors' calculations.*

Figure 2 shows the mechanism directly, plotting the ninety-five per cent predictive band of BNN for San Francisco against the realized path, alongside with BART's fixed per-window width. As we observe from the figure, the BNN band contracts through the quiet stretches — when the conditional variance is low and a narrow interval suffices — and widens through the crisis episodes marked by the shaded bands, when volatility spikes and a narrow interval would be pierced. BART's width, set once per estimation window,

cannot do this, and it is in exactly the crisis stretches, where the fixed width is far too narrow for the realized moves, that its breaches accumulate.

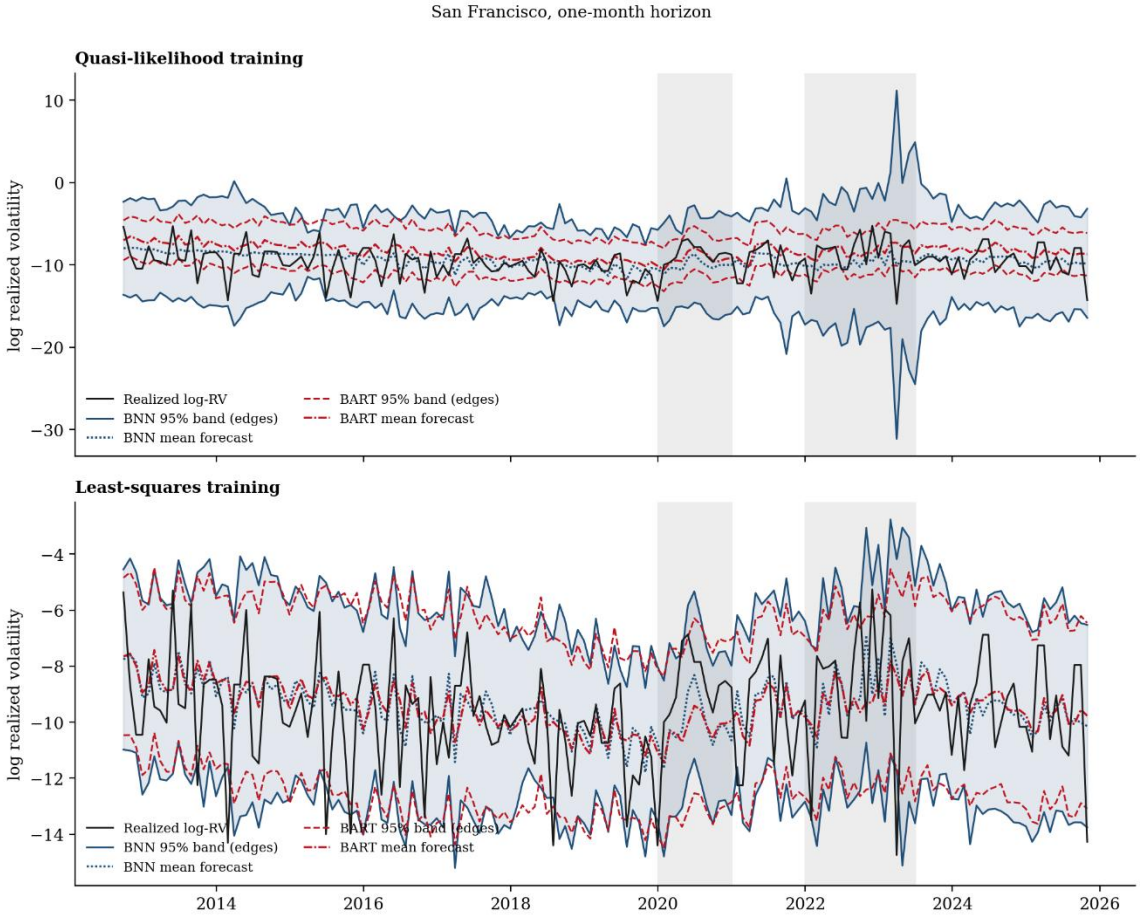


Figure 2. Predictive intervals of the BNN and BART against realized volatility (San Francisco, one-month horizon). *Upper panel quasi-likelihood training, lower panel least-squares training. The solid black line is realized log realized volatility, solid blue lines are the edges of the BNN's ninety-five per cent predictive interval, with its mean forecast dotted, dashed red lines are the edges of BART's interval, with its mean forecast dash-dotted. Shaded vertical bands mark the crisis episodes. The BNN's interval contracts in calm periods and widens through the crises, ranging from 5.3 to 42.3 log units, while BART's width is set once per estimation window and varies only from 4.7 to 6.4. Authors' calculations.*

Averaged over cities and windows, the filtered conditional scale ranges by 257 per cent peak-to-trough inside a single sixty-month estimation window under quasi-likelihood training and by 196 per cent under least-squares training, with within-window coefficients of variation of 0.285 and 0.220. These magnitudes are interpretable against the fixed-variance competitor, whose predictive scale over the evaluation sample carries coefficients of variation of 0.077 and 0.069 under the two objectives, a quarter of the

BNN's and attributable to the estimation window rolling forward rather than to any response within it. The score-driven filter is therefore reacting to the flow of shocks inside a fixed window, which is the motion a single scale per window cannot reproduce. The contribution of the score-driven variance is therefore not that it forecasts the level better than a disciplined tree, which it does not, but that it makes the tail controllable. That control is what capital applications price, since a supervisor and a capital rule respond to the exceedance count rather than to average accuracy. This is the empirical content, and the empirical limit, of the optimality result of Beutner, Lin and Lucas (2026), which establishes that a score-driven filter tracks the Kullback–Leibler-optimal parameter path, and that tracking governs the tail across the whole horizon range even where it earns nothing on the level.

3.4 Shrinkage and tail calibration

Two features of the fitted BNNs explain the calibration behavior documented above on how hard the prior shrinks, and how much the estimated variance moves. Figure 3 reports the intensity of shrinkage in the sense of the posterior slab variance of the regularized-horseshoe prior by feature set. Shrinkage tightens monotonically as predictors are added, with the slab variance falling from the autoregressive to the shock to the national feature sets, with the prior clamping down as the design matrix fills with uninformative columns, exactly as intended. And it is uniformly one and a half to two times stronger under quasi-likelihood training than under least-squares training, in every feature set. A more heavily shrunk mean model leaves more residual variance in the predictive density, which is why the quasi-likelihood intervals come out wide and its tails conservative, and why the lighter-shrinking least-squares objective produces the sharper, hotter-tailed densities of Tables 6 to 8. Together the two features fix the direction of miscalibration in each regime: harder shrinkage and a moving variance under quasi-likelihood give conservative tails, lighter shrinkage the reverse.

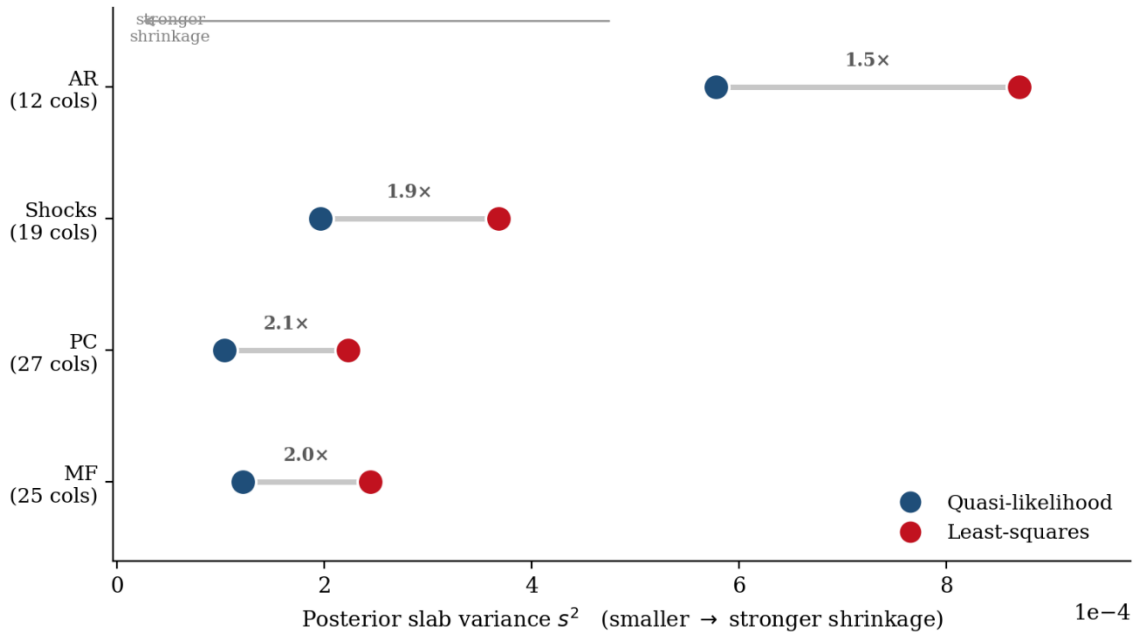


Figure 3. Intensity of shrinkage by feature set and training regime. *Posterior mean slab variance s^2 of the regularised-horseshoe prior by feature set, with column counts in parentheses, averaged across horizons, architectures and windows. Each pair of points joins the quasi-likelihood and least-squares estimates for one feature set; the connector label gives their ratio. Smaller values denote stronger shrinkage. Authors' calculations.*

The two training regimes differ not in how accurately they forecast the level of volatility, but in the direction their predictive densities miscalibrate. The quasi-likelihood objective shrinks the mean harder and drives its variance state through a score-driven filter that BART's fixed scale cannot match (Section 3.3), and a more heavily shrunk model with a moving variance leaves more residual dispersion in its predictive density. Figure 4 shows the consequence directly, through the probability integral transform of each model's one-month predictive densities, which should be uniform when the densities are correctly calibrated. Under quasi-likelihood training the histogram is hump-shaped, outcomes falling in the interior far more often than a calibrated density would place them, a signature of intervals that are too wide. Under least-squares training it is close to uniform through the interior but lifts at both extremes, a sign of intervals that are slightly too narrow, so that the tails are pierced more often than declared.

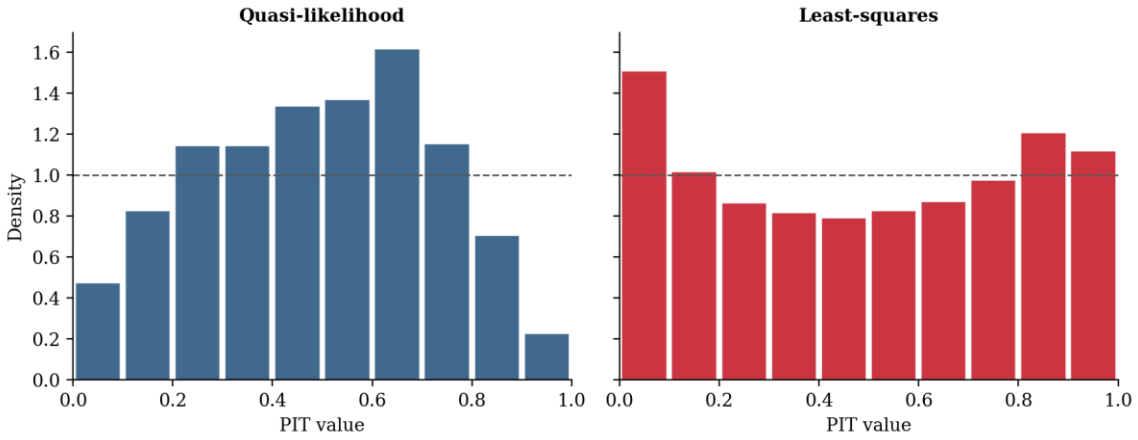


Figure 4. Calibration of the one-month predictive densities under the two training regimes. *Probability integral transform histograms for the leading BNN in each regime; a calibrated density gives a uniform histogram at the dashed line. A central hump indicates intervals that are too wide; raised extremes indicate intervals that are too narrow. Authors' calculations.*

Therefore, the least-squares regime is the better distributional forecaster in the ordinary sense, its density closer to uniform through the bulk while the quasi-likelihood regime is the one whose stated tail is the more believable, erring toward caution rather than overconfidence. The choice of training objective is thus not an internal technicality but a decision with a supervisory footprint, and reporting a single objective (Hauzenberger et al., 2025; Goulet Coulombe, Frenette and Klieber, 2026), conceals it. A model confidence set (Hansen, Lunde and Nason, 2011) confirms that at these sample sizes the field is largely inseparable on accuracy, with between seventeen and twenty-eight of the thirty-three models — the twenty-nine panel specifications and four BART benchmarks — surviving at every horizon and under both objectives (Appendix A3). The accuracy ranking has no reliable winner, and the analysis therefore turns to calibration, where the differences between the objectives are large.

The difference between the regimes is a difference in reported dispersion rather than in the location of the predictive density. Table A5 decomposes it. The two objectives forecast the conditional mean about equally well, with realized forecast-error standard deviations of 1.95 and 1.88 at one month and 2.33 and 2.61 at twelve, the least-squares mean marginally the better at the short horizon and the worse at the long one. What separates them is the dispersion each reports around that mean. The quasi-likelihood predictive standard deviation is 1.65 times the realized error dispersion at one month against 0.93 for least-squares, and the two ratios then move apart as the horizon lengthens, the quasi-likelihood excess easing to 1.47 while the least-squares deficiency deepens to 0.65. Rescaling both densities to the scale their own errors require removes the coverage gap altogether, turning +0.068, +0.096 and +0.140 at the three horizons into

−0.025, −0.027 and −0.026. The estimation loss therefore operates through a single channel, the width of the density reported around a mean forecast it barely affects, and the widening of the gap with the horizon is the widening of that channel rather than any new mechanism appearing at longer horizons. More details are reported in Appendix A.5.

The difference between the regimes is not a feature of the particular sample split. A stationary block bootstrap that resamples months in blocks while holding the full ten-city cross-section intact, so that both the national co-movement and the serial dependence of the panel are preserved, places the regime gap in ninety-five per cent coverage at 0.068 at one month, 0.096 at six and 0.140 at twelve, with bootstrap confidence intervals of [0.055, 0.080], [0.069, 0.129] and [0.095, 0.190] and p-values below 0.001 respectively. The corresponding gaps in the one per cent value-at-risk exceedance rate are −0.024, −0.037 and −0.051, on the same intervals and significance. The gap widens with the horizon, from 0.068 at one month to 0.096 at six and 0.140 at twelve[9].

3.5 Architecture and information

Two ablation tests establish that the density accuracy of the BNN is neither tunable through capacity nor improvable through information. The first concerns architecture. Table 9 reports the one-month density gain for every feature set across the four BNN configurations, in the manner of Hauzenberger et al. (2025, Table 2), under both training regimes. Within any feature set, the spread across the four architectures never exceeds 0.015 log points, an order of magnitude smaller than the differences between feature sets, and the Diebold–Mariano statistics confirm that almost no architecture is separable. Architecture does not matter, and it does not matter for a specific reason: the neuron-level scale of the regularized horseshoe fixes the effective width of the BNN from the data before the nominal grid can bind, so that a twenty-node network with switched-off nodes behaves as a smaller one. This vindicates the prior's design, extending the slab component of Piironen and Vehtari (2017) from coefficients to nodes, and it removes the principal practical objection to Bayesian neural networks in short samples, that they require a validation set the data cannot spare, without the architectural restrictions alternative approaches impose to the same end (Goulet Coulombe, Frenette and Klieber, 2026).

Table 9. Density gain by network architecture and feature set, one-month horizon

[9] The bootstrap uses an expected block length of twelve months. The conclusions are unchanged for expected block lengths of six, twenty-four and thirty-six months: the point estimates are invariant to the block length by construction, and the largest movement in any confidence-interval endpoint across the four settings is 0.015 coverage points, in the twelve-month-horizon coverage gap, whose interval runs from [0.104, 0.181] at the shortest block length to [0.089, 0.196] at the longest. Every interval excludes zero at every setting and every horizon.

Feature set	N10 L1	N10 L2	N20 L1	N20 L2	Spread	max |DM|
Panel A. Quasi-likelihood training						
AR	-0.041	-0.045	-0.041	-0.045	0.005	0.79
Shocks	-0.084	-0.070	-0.075	-0.069	0.015	1.91
PC	-0.031	-0.030	-0.030	-0.021	0.009	1.35
MF	-0.044	-0.047	-0.042	-0.037	0.010	1.42
Panel B. Least-squares training						
AR	+0.018	+0.020	+0.018	+0.018	0.002	1.25
Shocks	+0.004	+0.005	+0.005	+0.004	0.001	0.63
PC	-0.009	-0.010	-0.005	-0.007	0.005	2.17
MF	+0.016	+0.018	+0.023	+0.017	0.007	2.39

Note: ΔALPL against the pooled HAR baseline. N denotes hidden nodes, L hidden layers. "Spread" is the range across the four architectures within a feature set; "max\|DM\|" is the largest Diebold–Mariano statistic between the best architecture and the other three (values below 1.65 indicate no separation at the ten per cent level). Authors' calculations.

The second test concerns information. Table 10 reports the density gain by feature set and horizon. At one month horizon, the four feature sets differ by less than a tenth of a log point and none improves materially on the autoregressive base, with the realized-shock and national blocks adding essentially nothing to the density. The ordering rearranges with horizon rather than setting, the macroeconomic components leading at twelve months under quasi-likelihood training but collapsing under least-squares, but no feature block at any horizon buys more than a few hundredths of a log point over a pooled autoregression. The horizon-dependence is a property of data, not of one specification, and the overriding pattern is the weak second-moment implying that the information set matters far less than a reading of the earlier literature would suggest

Table 10. Density gain by feature set and horizon, leading architecture

Feature set	h=1	h=6	h=12
Panel A. Quasi-likelihood training			
AR	-0.041*	+0.009	+0.047
Shocks	-0.069***	+0.006	+0.048
PC	-0.021†	-0.003	+0.063†
MF	-0.037*	+0.015†	-0.051
Panel B. Least-squares training			
AR	+0.020	+0.021	-0.030
Shocks	+0.005	+0.023†	-0.025†
PC	-0.005*	-0.392**	-0.530**
MF	+0.023†	-0.019	-0.228

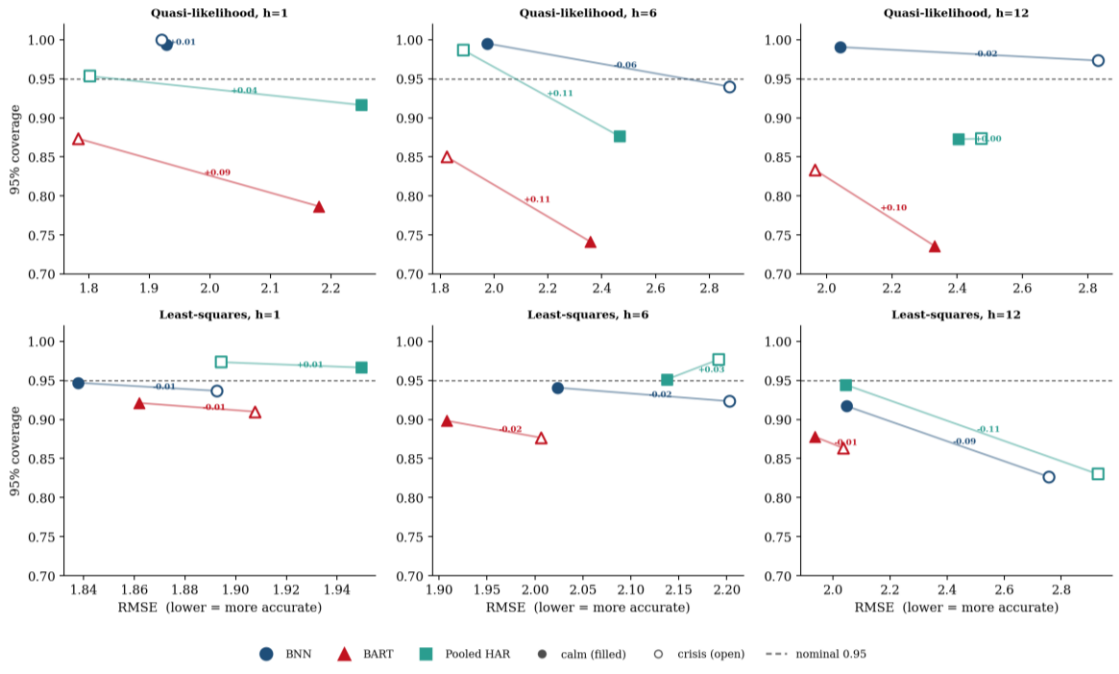
Note: best-architecture Δ ALPL against the pooled HAR baseline for each feature set at each horizon; † marks the leading feature set

within a column; Diebold–Mariano asterisks (at 10, 5, 1 per cent) test each feature set against that leader. Authors' calculations.

The two ablations together support that density accuracy does not respond to capacity and information, so it is not the margin on which the BNN earns its place. What does respond, and sharply, is calibration, and it responds to the one thing neither ablation varies, the training objective.

3.6 Performance in calm and crisis periods

An average over an eighteen-year sample can conceal a model that holds together only in ordinary conditions, which for a capital application is the case that matters least. We therefore split the sample into the crisis episodes — the COVID-19 shock of 2020 and the rate-hike period from January 2022 to June 2023[10] — and read the calibration of each training regime separately in the two. Figure 5 plots each model in the plane of accuracy (RMSE) and calibration (95% coverage) for the leading model of each family under both training regimes. A filled marker is the calm-period position, an open marker the crisis position, and the connector between them, labelled with the change in coverage, shows how each model's calibration moves under stress. The desirable region is the upper left, accurate and covering near the nominal 0.95.



[10] The sample begins in September 2007 and the rolling scheme uses a sixty-month training window plus the forecast horizon, so the first evaluated forecast falls in 2012. The 2007–09 financial crisis is therefore absorbed by the initial training window and does not enter the evaluation sample.

Figure 5. Accuracy against calibration in calm and crisis periods, by horizon and training regime. *Each marker places the leading model of a family, selected by average log predictive likelihood, in the plane of root mean squared error and ninety-five per cent coverage, computed separately over the crisis episodes (the 2020 COVID-19 shock and the rate-hike period from January 2022 to June 2023) and the remaining calm months. Circles are the BNN, triangles BART and squares the pooled HAR; filled markers are calm and open markers crisis, and the connector label gives the calm-to-crisis change in coverage. The dashed line marks the nominal 0.95, so vertical position measures calibration and horizontal position accuracy, with a model preferred on both criteria lying towards the upper left. The pooled HAR is the leading classical estimator by log predictive likelihood in every panel but one, where the static common-correlated-effects estimator leads. Authors' calculations.*

In calm conditions the leading quasi-likelihood BNN is alone in the upper left, holding the lowest root mean squared error in the figure together with coverage at or above nominal. Under stress no model keeps both, and the movements trace a trade-off rather than a ranking. The quasi-likelihood BNN retains its coverage — the calm-to-crisis change never exceeds six hundredths of a point — but its point accuracy deteriorates, its crisis root mean squared error rising to the highest in the figure at the six- and twelve-month horizons. This is the quasi-likelihood objective behaving as designed: by down-weighting extreme observations it declines to chase the crisis spikes, accepting a larger point error on those months in exchange for intervals that continue to contain them. BART makes the opposite exchange, improving its point accuracy in the crisis — at the longer horizons it posts the lowest crisis root mean squared error in the figure — while covering well below nominal throughout, so it holds accuracy at the cost of a tail that is pierced. The two models occupy opposite corners under stress, and which corner is preferable is not a question of general forecasting skill but of the loss the user faces, since a capital rule counts breaches rather than squared errors. Figure 5 reports the leading specification in each family; Table A2 in the Appendix gives the calm and crisis coverage of every model at this horizon, ordered by distance from nominal, and shows the same pattern across the full field.

The pooled HAR's coverage is near nominal on average, which on a single-period reading would make it look well calibrated, but its connectors are among the longest in the figure, swinging by +0.11 from calm to crisis at six months and by –0.11 at twelve. A coverage that depends this strongly on the regime is not dependable, and its favorable average conceals a calibration that a capital rule applied period by period would not find reliable. Stability across regimes, not the average level alone, is what a calibration must have to be useful under stress, and it is the short connectors of the quasi-likelihood BNN, not the position of the pooled HAR, that display it.

The contrast between the two BNN regimes is the same one that runs throughout the paper, now measured under stress. Under least-squares training the BNN's coverage is already below nominal in calm months and falls further in crisis, most sharply at twelve months, where the change is −0.09; under quasi-likelihood training it holds. In supervisory units the one per cent value-at-risk exceedances in the crisis at six months are 4.4 per 250 observations under quasi-likelihood training against 21.1 under least-squares, the former at the edge of the Basel green zone and the latter deep in the red. The zone is what sets capital. A green record carries the base multiplier of three, a red record carries four (Basel Committee on Banking Supervision, 1996), so two estimations of the same model on the same data, differing only in the loss minimized during training, imply a third more required capital on an identical book of exposures, with a red-zone record liable to suspend the internal model altogether. The mapping from a monthly panel to a 250-observation supervisory window is a stylization and the multiplier is therefore indicative, but the zone crossing that determines whether capital is added is not sensitive to it. Table A4 reports the full backtesting record at the one-month horizon, giving the coverage, breach count, zone and multiplier for the leading model of each family under both objectives, in calm months and in crisis. The mechanism is the one documented in Sections 3.3 and 3.4, and it is not that one regime becomes fragile in a crisis while the other does not, since both filter the same variance state through the same episodes. It is that the least-squares density is already the sharper and hotter-tailed of the two in calm conditions, so a crisis pierces its intervals first, while the conservatively wide quasi-likelihood intervals retain the margin to absorb the same moves.

3.7 Cross-city robustness of the calibration result

The results so far are averages over the ten markets. A panel estimator is only as convincing as its behavior in the cities it pools, so we ask whether the central result of the paper — that the training objective governs tail calibration — holds market by market, and whether its size varies with the character of the market. Figure 6 reports the ninety-five per cent coverage of the leading BNN under each objective for all ten markets, with the size of each bubble giving the number of one per cent value-at-risk breaches per 250 observations. In every market, the quasi-likelihood markers sit above the nominal 0.95 and carry small bubbles, while the least-squares markers sit below nominal and carry larger ones. In fact, the least-squares bubbles grow as the horizon lengthens from one to twelve months. The ordering never reverses: in every one of the ten markets quasi-likelihood coverage exceeds least-squares coverage, so the pooled result is not the product of a few markets but a property of the estimation loss across the panel.

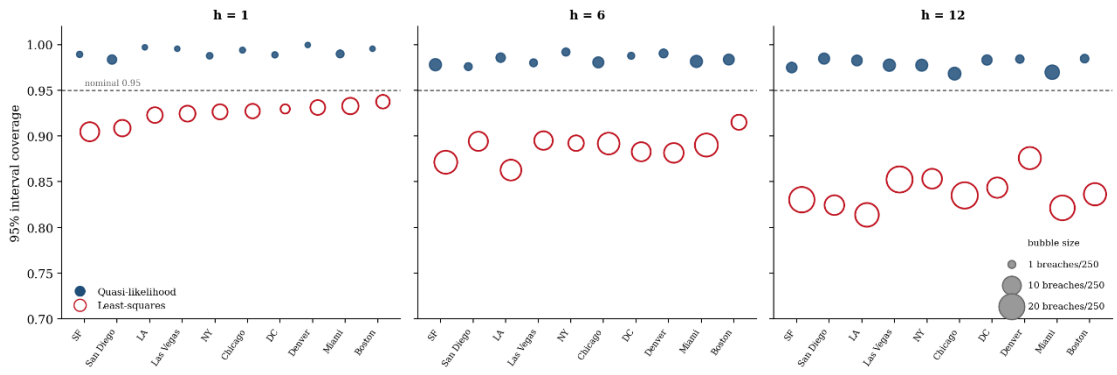


Figure 6. Coverage and tail breaches by city, horizon and training objective. *Ninety-five per cent interval coverage of the leading BNN under quasi-likelihood (filled) and least-squares (open) training, by city, at the one-, six- and twelve-month horizons, with the dashed line at the nominal 0.95. Bubble area is the number of one per cent value-at-risk breaches per 250 observations, so a desirable model sits high and small. Quasi-likelihood coverage lies above nominal with few breaches in every market; least-squares coverage lies below nominal, with breaches that grow as the horizon lengthens. Authors' calculations.*

Figure 7 plots each market's coverage under the two training objectives against the level of its realized volatility. Quasi-likelihood coverage is flat in the volatility level, with a correlation of essentially zero: the conservative intervals hold above nominal in the calmest and the most volatile markets alike. Least-squares coverage falls with the volatility level, at a correlation of –0.68, dropping from near nominal in the least volatile market, Washington DC, toward the region where breaches accumulate in the most volatile, San Francisco. The gap between the objectives therefore widens in volatile markets not because the quasi-likelihood objective improves but because the least-squares objective deteriorates, and it deteriorates precisely where there is most volatility to track; the markets a risk manager is most concerned to calibrate.

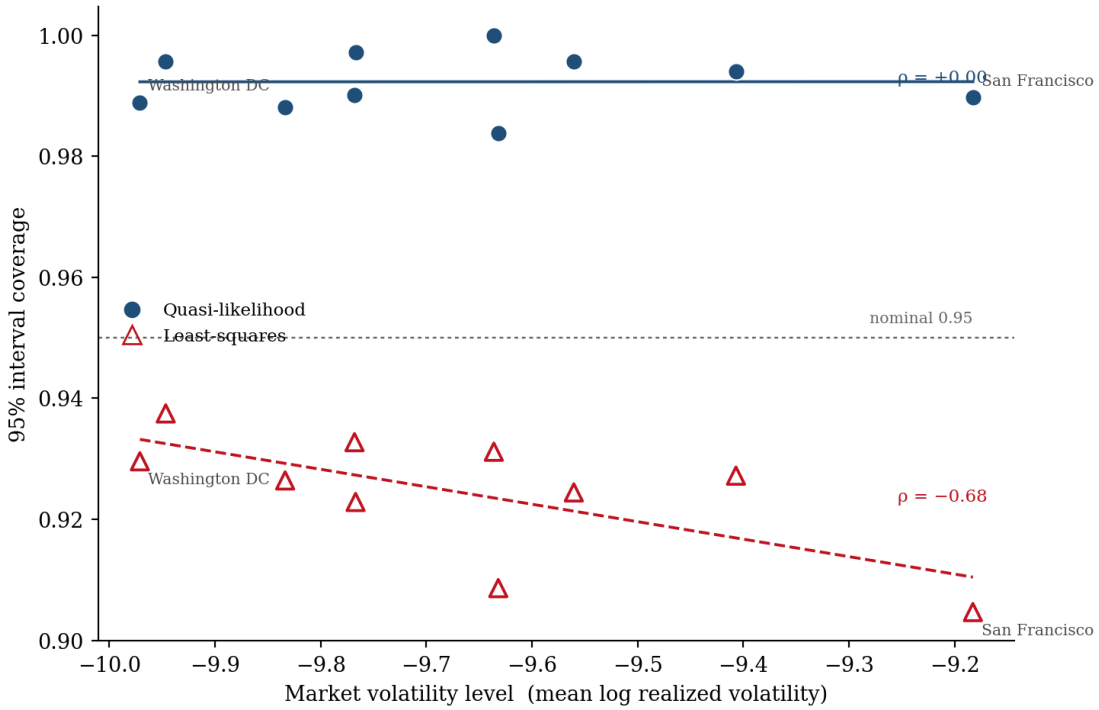


Figure 7. Coverage against market volatility, by training objective. *One-month ninety-five per cent coverage of the leading BNN under each objective, by city, against the market's mean log realized volatility; dashed lines are least-squares fits annotated with the correlation. Quasi-likelihood coverage is flat in the volatility level, least-squares coverage falls, so the gap between them widens in more volatile markets. Authors' calculations.*

What the coverage gap follows is the amount of volatility a market carries, and not the other characteristics one might expect. It does not track the kurtosis of the within-month return distribution, at a correlation of −0.11, and it does not track housing supply elasticity: arranging the ten markets by the Saiz (2010) elasticity and splitting at the median — Miami, Los Angeles, San Francisco, San Diego and New York in the inelastic group, Chicago, Boston, Las Vegas, Denver and Washington DC in the elastic — leaves no difference in the gap between the two groups. Supply elasticity governs the size and persistence of housing price booms (Glaeser, Gyourko and Saiz, 2008; Aastveit, Albuquerque and Anundsen, 2023), but it does not govern how much the training objective matters for calibration. Thus, the markets in which the least-squares buffer is least reliably sized are the volatile ones, and these are not the supply-constrained markets that draw the most regulatory attention — volatility level and supply elasticity are essentially uncorrelated across the ten markets, the most volatile of them inelastic but the next two elastic — so the cities where a risk buffer most needs care are not the cities a supply-based reading would flag. With ten markets the relationship is descriptive, and is reported as such.

4. Conclusion

Two estimations of the same model, on the same data, differing only in the loss minimized during training, produce predictive densities that err in opposite directions at the tail. That is the central finding of this paper, and it is one with a price. Because supervisory capital is set by counting the occasions on which a model's stated quantile is breached, the choice between a proxy-robust loss and a likelihood — a decision internal to a modelling team and invisible to anyone reading the forecasts — moves an internal model between capital zones. We reach that result through a panel Bayesian neural network with a regularized-horseshoe prior and a city-specific score-driven exponential GARCH variance equation, estimated under both objectives and set against the full classical panel toolkit and a tree ensemble across ten metropolitan housing futures markets over eighteen years.

The results confirm the view received on two counts and contradict it on a third. They confirm, that the predictability of the second moment is weak. No model, classical or neural, improves on a pooled autoregression of the realized measure by more than about six-hundredths of a log point at any horizon under either objective, and the model confidence set retains between seventeen and twenty-eight of the thirty-three specifications compared, so the accuracy ranking has no reliable winner. This is the panel counterpart of the bound the realized-volatility literature has documented for single series, and the interval-based work of Bollerslev, Patton and Quaedvlieg (2016) anticipates the reason, in that the proxy is noisiest in exactly the turbulent months whose tails a forecast must capture. Moreover, our results confirm, the central message of the panel forecasting literature, that no single pooling scheme dominates and that the return to pooling depends on the structure of heterogeneity and the time-series length available (Pesaran, Pick and Timmermann, 2026). Our classical benchmarks reproduce that pattern, with the pooled and common-correlated-effects estimators trading places across horizons and the mean-group and dynamic variants deteriorating as parameters proliferate.

Where the results depart from the received view is on what a flexible model contributes. The presumption in the machine-learning forecasting literature is that nonlinear models earn their advantage through a richer conditional mean. Here they do not. A tree ensemble carrying the same features and the same nonlinear mean, but a single scale per estimation window, matches the BNN at one month under a proper likelihood and beats it at longer ones, so the score-driven variance equation buys no advantage in point or density accuracy. Its value lies in the tail, where it makes the level of miscalibration controllable, and this is the mechanism through which the paper's central result operates.

That result concerns the training objective. The loss minimized in estimation determines not the size but the direction of a model's miscalibration. Quasi-likelihood training yields conservative, over-wide intervals with coverage near 0.99 and tail exceedances well inside nominal, while likelihood training yields sharper densities with overconfident tails,

coverage near 0.94 and exceedances above nominal. These are the same network on the same data, separated only by the estimation loss, and a panel block bootstrap places the coverage gap at 0.068, 0.096 and 0.140 at the one-, six- and twelve-month horizons with p-values below 0.001, with the gap widening with the horizon. That the objective optimized shapes what a forecast is good at is the founding insight of the scoring-rule literature (Gneiting and Raftery, 2007; Nolde and Ziegel, 2017). What that literature has framed as a question about evaluation — which score should rank competing procedures — we answer on the estimation side, and with a sign attached: the choice of loss fixes which of two economically opposite errors the user inherits. The transmission runs through the variance equation, the component that earns nothing on average accuracy, so the two findings are one. A fixed-variance competitor has no lever through which the choice of loss can reach the tail.

The economic consequence follows from the arithmetic of supervision rather than from any further modelling. Under the Basel traffic-light test an internal model is judged by counting exceptions to its ninety-nine per cent value-at-risk forecast over 250 observations, with four or fewer placing it in the green zone at the base multiplier, five to nine in the yellow at graduated steps and ten or more in the red (Basel Committee on Banking Supervision, 1996). Our two regimes are identical in architecture, data and features and coincide in the green zone in calm conditions, but part under stress, the quasi-likelihood model holding green while the equally accurate likelihood-trained model crosses into the penalty zones. The estimation loss, normally treated as an internal choice orthogonal to supervision, is therefore neither internal nor orthogonal. A firm that selects its estimator on in-sample fit or on average density accuracy may select into a capital surcharge, and a model validation conducted on density scores can rank specifications in the opposite order to one conducted on exception counts. This is the applied content of the incentive argument made theoretically by Nolde and Ziegel (2017), that the function used to judge a risk-measurement procedure shapes the procedures that get built.

Three limitations bound these claims and point in one direction. The evaluation rests on realized coverage and exceedance counts rather than on the weighted or censored scoring rules developed for tail evaluation (Diks, Panchenko and van Dijk, 2011; Mitchell and Weale, 2023; Allen et al., 2025), a choice made to keep the units those in which the supervisory decision is taken, at some cost in statistical efficiency, and a formal tail-calibration diagnostic applied to a panel of this kind is a natural next step that would require extending those tools to a cross-section with a common factor. The crisis evidence rests on a short window. The rolling scheme places the first evaluated forecast in 2012, so the 2007–09 collapse — the episode a housing-volatility model would most want to be judged against — lies inside the initial estimation window and never enters the evaluation, and the calm-versus-crisis contrast is therefore identified from thirty months spanning a pandemic and a monetary tightening rather than a housing-led downturn. The extreme quantiles stay beyond reach for every specification. A sixty-month estimation

window identifies the center of a predictive density far better than its 99.5th percentile, and no estimator in the comparison escapes that constraint, so the tail region where a stress calibration is set is precisely the region the data identify least well.

That the advantage the network delivers flows through its variance process rather than its mean identifies where the next gain should be sought. A variance process made multivariate across cities, so that the crisis-period widening documented above is shared and propagated rather than filtered market by market, would address the tail limitation and the diversification question at once, the latter bearing directly on the premise of geographic diversification underlying credit-risk-transfer structures. Whether the calibration asymmetry documented here is specific to volatility, where the target is a noisy proxy, or extends to any forecasting problem in which a proxy-robust loss competes with a likelihood, is the broader question this paper leaves open.

References

Aastveit, K.A., Albuquerque, B. and Anundsen, A.K. (2023) 'Changing supply elasticities and regional housing booms', *Journal of Money, Credit and Banking*, 55(7), pp. 1749-1783.

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

Andersen, T.G. and Bollerslev, T. (1998) 'Answering the skeptics: yes, standard volatility models do provide accurate forecasts', *International Economic Review*, 39(4), pp. 885-905.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P. (2003) 'Modeling and forecasting realized volatility', *Econometrica*, 71(2), pp. 579-625.

Aquaro, M., Bailey, N. and Pesaran, M.H. (2021) 'Estimation and inference for spatial models with heterogeneous coefficients: an application to US house prices', *Journal of Applied Econometrics*, 36(1), pp. 18-44.

Barndorff-Nielsen, O.E. and Shephard, N. (2002) 'Econometric analysis of realized volatility and its use in estimating stochastic volatility models', *Journal of the Royal Statistical Society: Series B*, 64(2), pp. 253-280.

Basel Committee on Banking Supervision (1996) *Supervisory framework for the use of '**backtesting**' in conjunction with the internal models approach to market risk capital requirements*. Basel: Bank for International Settlements.

Beutner, E., Lin, Y. and Lucas, A. (2026) 'Consistency, distributional convergence, and optimality of time-varying parameters in score-driven models', *Journal of Econometrics*, 255, 106218.

Bollerslev, T., Patton, A.J. and Quaedvlieg, R. (2016) 'Exploiting the errors: a simple approach for improved volatility forecasting', *Journal of Econometrics*, 192(1), pp. 1-18.

Bollerslev, T., Patton, A.J. and Wang, W. (2016) 'Daily house price indices: construction, modeling, and longer-run predictions', *Journal of Applied Econometrics*, 31(6), pp. 1005-1025.

Camehl, A. (2023) 'Penalized estimation of panel vector autoregressive models: a panel LASSO approach', *International Journal of Forecasting*, 39(3), pp. 1185-1204.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010) 'The horseshoe estimator for sparse signals', *Biometrika*, 97(2), pp. 465-480.

Cheng, C.H.J. and Chiu, C.-W. (2020) 'Nonlinear effects of mortgage spreads over the business cycle', *Journal of Money, Credit and Banking*, 52(6), pp. 1593-1620.

Chipman, H.A., George, E.I. and McCulloch, R.E. (2010) 'BART: Bayesian additive regression trees', *Annals of Applied Statistics*, 4(1), pp. 266-298.

Chudik, A. and Pesaran, M.H. (2015) 'Common correlated effects estimation of heterogeneous dynamic panel data models with weakly exogenous regressors', *Journal of Econometrics*, 188(2), pp. 393-420.

Corsi, F. (2009) 'A simple approximate long-memory model of realized volatility', *Journal of Financial Econometrics*, 7(2), pp. 174-196.

Creal, D., Koopman, S.J. and Lucas, A. (2013) 'Generalized autoregressive score models with applications', *Journal of Applied Econometrics*, 28(5), pp. 777-795.

Del Negro, M. and Otrok, C. (2007) '99 Luftballons: monetary policy and the house price boom across U.S. states', *Journal of Monetary Economics*, 54(7), pp. 1962-1985.

Di Maggio, M. and Kermani, A. (2017) 'Credit-induced boom and bust', *Review of Financial Studies*, 30(11), pp. 3711-3758.

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

Diks, C., Panchenko, V. and van Dijk, D. (2011) 'Likelihood-based scoring rules for comparing density forecasts in tails', *Journal of Econometrics*, 163(2), pp. 215–230.

Fairchild, J., Ma, J. and Wu, S. (2015) 'Understanding housing market volatility', *Journal of Money, Credit and Banking*, 47(7), pp. 1309-1337.

Favara, G. and Imbs, J. (2015) 'Credit supply and the price of housing', *American Economic Review*, 105(3), pp. 958-992.

Federal Housing Finance Agency (2020) 'Enterprise Regulatory Capital Framework', Final Rule, 85 Fed. Reg. 82150, 17 December.

Federal Housing Finance Agency (2024) *Credit Risk Transfer Progress Report: Fourth Quarter 2023*. Washington, DC: FHFA.

Glaeser, E.L., Gyourko, J. and Saiz, A. (2008) 'Housing supply and housing bubbles', *Journal of Urban Economics*, 64(2), pp. 198-217.

Goulet Coulombe, P., Frenette, M. and Klieber, K. (2026) 'From reactive to proactive volatility modeling with hemisphere neural networks', *Journal of Applied Econometrics*, 41(3), pp. 265-279.

Greenwald, D.L. and Guren, A. (2025) 'Do credit conditions move house prices?', *American Economic Review*, 115(10), pp. 3559-3596.

Gupta, R., Ma, J., Theodoridis, K. and Wohar, M.E. (2023) 'Is there a national housing market bubble brewing in the United States?', *Macroeconomic Dynamics*, 27(8), pp. 2191-2228.

Hansen, P.R. and Lunde, A. (2005) 'A forecast comparison of volatility models: does anything beat a GARCH(1,1)?', *Journal of Applied Econometrics*, 20(7), pp. 873-889.

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

Harvey, A.C. (2013) *Dynamic models for volatility and heavy tails*. Cambridge: Cambridge University Press.

Harvey, A.C. and Chakravarty, T. (2008) *Beta-t-(E)GARCH*. Cambridge Working Papers in Economics 0840, University of Cambridge.

Hauzenberger, N., Huber, F., Klieber, K. and Marcellino, M. (2025) 'Bayesian neural networks for macroeconomic analysis', *Journal of Econometrics*, 249, 105843.

Jurado, K., Ludvigson, S.C. and Ng, S. (2015) 'Measuring uncertainty', *American Economic Review*, 105(3), pp. 1177-1216.

Justiniano, A., Primiceri, G.E. and Tambalotti, A. (2019) 'Credit supply and the housing boom', *Journal of Political Economy*, 127(3), pp. 1317-1350.

Karoglou, M., Morley, B. and Thomas, D. (2013) 'Risk and structural instability in US house prices', *Journal of Real Estate Finance and Economics*, 46(3), pp. 424-436.

Kleen, O. (2024) 'Scaling and measurement error sensitivity of scoring rules for distribution forecasts', *Journal of Applied Econometrics*, 39(5), pp. 833–849.

Liu, L., Moon, H.R. and Schorfheide, F. (2020) 'Forecasting with dynamic panel data models', *Econometrica*, 88(1), pp. 171-201.

llen, S., Koh, J., Segers, J. and Ziegel, J. (2025) 'Tail calibration of probabilistic forecasts', *Journal of the American Statistical Association*, 120(552), pp. 2796–2808.

Loutskina, E. and Strahan, P.E. (2015) 'Financial integration, housing, and economic volatility', *Journal of Financial Economics*, 115(1), pp. 25-41.

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

Mian, A. and Sufi, A. (2022) 'Credit supply and housing speculation', *Review of Financial Studies*, 35(2), pp. 680-719.

Miller, N. and Peng, L. (2006) 'Exploring metropolitan housing price volatility', *Journal of Real Estate Finance and Economics*, 33(1), pp. 5-18.

Mitchell, J. and Weale, M. (2023) 'Censored density forecasts: production and evaluation', *Journal of Applied Econometrics*, 38(5), pp. 714–734.

Møller, S.V., Pedersen, T.Q., Montes Schütte, E.C. and Timmermann, A. (2024) 'Search and predictability of prices in the housing market', *Management Science*, 70(1), pp. 415-438.

Nazlioglu, S., Tidwell, A. and Lee, J. (2026) 'Heterogeneous co-movements in US state and metropolitan statistical area housing prices: new insights from quantile factor models', *Real Estate Economics*, 54(4), pp. 1080-1104.

Nelson, D.B. (1991) 'Conditional heteroskedasticity in asset returns: a new approach', *Econometrica*, 59(2), pp. 347-370.

Nolde, N. and Ziegel, J.F. (2017) 'Elicitability and backtesting: perspectives for banking regulation', *Annals of Applied Statistics*, 11(4), pp. 1833–1874.

Patton, A.J. (2011) 'Volatility forecast comparison using imperfect volatility proxies', *Journal of Econometrics*, 160(1), pp. 246-256.

Pesaran, M.H. (2006) 'Estimation and inference in large heterogeneous panels with a multifactor error structure', *Econometrica*, 74(4), pp. 967-1012.

Pesaran, M.H. and Smith, R. (1995) 'Estimating long-run relationships from dynamic heterogeneous panels', *Journal of Econometrics*, 68(1), pp. 79-113.

Pesaran, M.H., Pick, A. and Timmermann, A. (2026) ‘Forecasting with panel data: estimation uncertainty versus parameter heterogeneity’, *Quantitative Economics*, 17(2), pp. 342-393.

Pesaran, M.H., Shin, Y. and Smith, R.P. (1999) ‘Pooled mean group estimation of dynamic heterogeneous panels’, *Journal of the American Statistical Association*, 94(446), pp. 621-634.

Piironen, J. and Vehtari, A. (2017) ‘Sparsity information and regularization in the horseshoe and other shrinkage priors’, *Electronic Journal of Statistics*, 11(2), pp. 5018-5051.

Plakandaras, V., Gupta, R., Katrakilidis, C. and Wohar, M.E. (2020) ‘Time-varying role of macroeconomic shocks on house prices in the US and UK: evidence from over 150 years of data’, *Empirical Economics*, 58(5), pp. 2249-2285.

Politis, D.N. and Romano, J.P. (1994) 'The stationary bootstrap', *Journal of the American Statistical Association*, 89(428), pp. 1303–1313.

Saiz, A. (2010) ‘The geographic determinants of housing supply’, *Quarterly Journal of Economics*, 125(3), pp. 1253-1296.

Tidwell, A., Lu, Y., Lee, J. and Banerjee, P. (2023) ‘Nature of comovements in US state and MSA housing prices’, *Real Estate Economics*, 51(4), pp. 959-989.

Umlandt, D. (2023) ‘Score-driven asset pricing: predicting time-varying risk premia based on cross-sectional model performance’, *Journal of Econometrics*, 237(2), 105470.