# Model Selection for Probabilistic Long-Horizon Portfolio Forecasting


## Abstract

This paper evaluates probabilistic models for multi-horizon portfolio-return forecasting using a rolling-origin out-of-sample design. The empirical panel contains 80 multi-asset portfolios with common daily history from December 31, 1979 through May 13, 2026. The documented model-selection record contains 211 scored forecasting specifications evaluated on the common 80-portfolio, 15-horizon rolling-origin design. Adjusted continuous ranked probability score (adjusted CRPS) denotes the fixed CRPS-style weighted-interval score defined in the paper. Forecasts are evaluated as full predictive distributions over horizons from 21 to 2,268 trading days.

The selected production model combines four elements: a latent expected-return process anchored to each portfolio's historical annualized log-growth rate, posterior-weighted integrated nested Laplace quadrature for stochastic-volatility hyperparameters and the terminal latent state, a history-gated calendar-geometric decomposition of latent log volatility with origin-estimated shared-state residual dynamics, and stationary-bootstrap resampling of mean-standardized empirical portfolio-return innovations. On the direct 15 fixed-horizon benchmark comparison, its adjusted CRPS is 0.229988, a 12.90% reduction relative to the naive benchmark. The portfolio-level paired loss summary gives $t = 8.06$ and a one-sided p-value below .001 under exchangeability of portfolio labels. Because the panel reuses constituent assets across portfolios, that calculation is conditional evidence for the locked panel rather than inference from 80 independent economic units. The production model is better than the naive benchmark at all 15 fixed horizons from 1 month through 9 years.

## 1. Introduction

Long-horizon portfolio planning requires a predictive distribution, not only an expected return. A useful forecasting model must represent mean uncertainty, volatility clustering, non-Gaussian residual behavior, and horizon-dependent dispersion. In practice, many plausible models can look reasonable in sample while producing materially different simulated distributions out of sample.

This paper studies the model-selection problem directly. For each portfolio, origin date, and investment horizon, a candidate model is fit only on information available at the origin and then used to simulate future cumulative returns. The realized future return is compared with the model's predictive distribution using the adjusted-CRPS score defined below: a fixed CRPS-style weighted-interval scoring criterion for probabilistic forecasts. Lower adjusted CRPS indicates a predictive distribution that is better calibrated and sharper around the realized outcome under that criterion.

The contribution is empirical and methodological. First, the paper documents a rolling-origin forecast-selection framework across a broad multi-asset portfolio panel. Second, it compares 211 distributional forecasting specifications under a common historical window and horizon grid. Third, it specifies a history-gated calendar-geometric, mode-centered, posterior-weighted integrated nested Laplace quadrature multiscale stochastic-volatility model whose future multiscale and residual-state dynamics are estimated from information available at each forecast origin. Fourth, it places that production specification against representative popularized Monte Carlo methodology analogs so that the comparison includes the main simulation families commonly used for long-horizon portfolio planning.

## 2. Related Literature

The evaluation design follows the forecasting literature's preference for out-of-sample and rolling-origin tests when comparing models intended for future use (Tashman, 2000). This is especially important for long-horizon portfolio simulation because in-sample likelihood or fit can favor models that reproduce historical volatility episodes without producing better predictive distributions at future origins.

The scoring criterion is grounded in the literature on proper scoring rules for probabilistic forecasts. Gneiting and Raftery (2007) show why proper scoring rules are needed when the object being evaluated is a full predictive distribution, and Gneiting and Ranjan (2011) discuss CRPS-based density forecast comparison. Bracher et al. (2021) establish the weighted interval score as a proper score for reported predictive quantiles and show its relationship to CRPS. In this paper, the reported adjusted-CRPS field is the fixed weighted interval score defined below, applied to the simulated cumulative-return distribution at each horizon.

The compared model set draws on the standard conditional-volatility lineage beginning with autoregressive conditional heteroskedasticity (ARCH; Engle, 1982), generalized autoregressive conditional heteroskedasticity (GARCH; Bollerslev, 1986), and asymmetric generalized autoregressive conditional heteroskedasticity volatility dynamics (Glosten, Jagannathan, and Runkle, 1993). The selected model is closer in spirit to stochastic-volatility models, where volatility is a latent state rather than a deterministic function of lagged squared residuals. Kim, Shephard, and Chib (1998) provide a standard likelihood and simulation-based treatment of stochastic-volatility models.

Additional methodological ingredients matter for the selected model. The return simulation uses the stationary bootstrap of Politis and Romano (1994), with automatic block-length logic connected to Politis and White (2004) and its correction by Patton, Politis, and White (2009). The expected-return and volatility states use linear-Gaussian filtering and smoothing logic associated with Kalman (1960) and Rauch, Tung, and Striebel (1965). Stochastic-volatility state inference uses deterministic integrated nested Laplace approximation logic: transformed posterior-mode optimization, local-Hessian scaling, posterior-corrected Gauss-Hermite integration over three state-space hyperparameters, and Gaussian quadrature over the terminal latent log-variance marginal. This construction connects the implementation to Tierney and Kadane (1986), Rue, Martino, and Chopin (2009), and the stochastic-volatility application of Martino et al. (2011).

## 3. Portfolio Panel

The evaluation panel contains 80 multi-asset portfolios. Each portfolio is observed on a common daily return history beginning December 31, 1979 and ending May 13, 2026, approximately 46.37 years. The portfolio-construction record identifies the panel as an equal-asset-class-history panel rebuilt from documented tickers, weights, and rebalancing rules.

The portfolio set is intentionally heterogeneous while preserving a common evaluation window. The portfolio appendix records 20 monthly-rebalance portfolios, 18 quarterly-rebalance portfolios, 18 annual-rebalance portfolios, and 24 buy-and-hold or unrebalanced portfolios. The average asset-class allocation is one-third equity, one-third fixed income, and one-third alternatives. Appendix A lists each portfolio’s holdings, target weights, and rebalancing convention.

Daily asset inputs enter the portfolio construction as simple returns. Effective portfolio weights therefore drift with asset performance between scheduled rebalances and are reset to the recorded target weights on a scheduled rebalance. At such a rebalance, one-way turnover tau is one-half the sum of the absolute differences between target and post-return drifted weights; tau is zero otherwise. The fixed panel construction charges c = 0.0015, or 15 basis points, per unit of one-way turnover. The resulting equity curve is converted first to net portfolio simple returns and only then, through the log-one-plus transformation, to the daily portfolio log returns used by every candidate and benchmark model.

$$R_{p,t}^{gross} = \sum_{i=1}^{N} w_{i,t-1} R_{i,t}, V_{p,t} = V_{p,t-1} \left(1 + R_{p,t}^{gross}\right)\left(1 - c\,\tau_t\right), r_{p,t} = \log\left(\frac{V_{p,t}}{V_{p,t-1}}\right) = \log\left(1 + R_{p,t}^{net}\right)$$

## 4. Rolling-Origin Forecast Design

For each portfolio, origin date, model specification, and horizon, the target is cumulative future log return:

$$Y_{p,o,h} = \sum_{j=1}^{h} r_{p,o+j}$$

At each origin, the candidate model is estimated using only data observed through that date. It then generates a Monte Carlo sample from the predictive distribution of the cumulative-return target:

$$\widehat{F}_{m,p,o,h} = \left\{ Y_{m,p,o,h}^{(s)} \right\}_{s=1}^{S}$$

The direct selected-model benchmark grid contains 15 fixed horizons:

$$H = \{21, 42, 63, 126, 189, 252, 378, 504, 756, 1008, 1260, 1512, 1764, 2016, 2268\}$$

| **Horizon group** | **Trading days** |
|---|---|
| Short and intermediate fixed horizons | 21, 42, 63, 126, 189, 252 |
| Long fixed horizons | 378, 504, 756, 1008, 1260, 1512, 1764, 2016, 2268 |

The primary full-panel fixed-horizon run for the selected model used 33 origins per portfolio, selected with an evenly spaced full-history origin policy, and 300 simulations per origin and horizon. A portfolio-origin-horizon cell is scored only when the realized terminal return is observed within the historical panel. This matters at long horizons: later origins can be used for short-horizon forecasts, but not for horizons that extend beyond the sample end. The fixed-horizon comparison produces 329 scored cells per portfolio, or 26,320 scored cells across the 80-portfolio panel.

The horizon-specific score counts for that direct comparison are:

| Horizon days | Scored cells |
|---|---|
| 21 | 2,640 |
| 42 | 2,560 |
| 63 | 2,560 |
| 126 | 2,560 |
| 189 | 2,480 |
| 252 | 2,400 |
| 378 | 2,320 |
| 504 | 2,080 |
| 756 | 1,760 |
| 1008 | 1,520 |
| 1260 | 1,280 |
| 1512 | 960 |
| 1764 | 720 |
| 2016 | 400 |
| 2268 | 80 |

The final production-model benchmark evidence uses the same saved 80-portfolio panel, the same 300-simulation design, and the same fixed 15-horizon grid. The documented comparison record combines the canonical 200-specification file with eleven later same-panel state-inference, stochastic-volatility, and multiscale-memory specifications, for 211 scored specifications in total.

The documented model-selection file contains 211 scored forecasting specifications. Each public row represents a distinct describable specification evaluated under the common-panel rule. The set includes the production posterior-weighted integrated nested Laplace quadrature specification, the naive benchmark, and representative alternatives spanning expected-return models, volatility processes, innovation distributions, particle-based state-inference methods, multiscale memory bases, and empirical resampling rules. Source-file provenance, origin count, simulation count, and panel-construction metadata are retained in the replication materials so that score rows can be traced to their generating experiments without exposing internal run labels in the manuscript.

For external reporting, the final model-selection set is the reconciled set of describable model specifications with a common-panel score. When repeated research runs exist for the same public specification, the appendix retains one row for that specification and preserves source provenance in the replication mapping. The ranking therefore compares model specifications under a common panel rather than transient run labels.

This design should be read as a pseudo-real-time forecasting exercise. For each origin, portfolio construction and model estimation use only information available at that origin. The full historical panel defines the common evaluation sample; it does not allow a model to condition on future returns when producing a forecast.

## 5. Scoring Rule

The scoring criterion is motivated by continuous ranked probability score. For a predictive distribution and realized outcome, analytic CRPS is

$$CRPS(F,y)=\int_{-\infty}^{\infty} \text{¿} F(z)-1(y\le z)\}^{2}\,dz\,\text{¿}$$

where F is the predictive cumulative distribution function and y is the realized outcome. With a finite Monte Carlo sample, the empirical score is:

$$\widehat{CRPS}=\frac{1}{S}\sum_{s=1}^{S}|Y^{(s)}-y|-\frac{1}{2S^{2}}\sum_{s=1}^{S}\sum_{q=1}^{S}|Y^{(s)}-Y^{(q)}|$$

CRPS is attractive here because it evaluates the entire predictive distribution. A model can improve the score by centering forecasts more accurately, by producing better-calibrated dispersion, or by avoiding distributional shapes that place too much probability mass far from realized outcomes.

The scoring harness reports an adjusted-CRPS field implemented as a finite-quantile weighted interval score approximation to CRPS on terminal log returns. The paper uses the field name adjusted CRPS because that is the implementation field name, but the implemented object is the fixed approximation below rather than the analytic CRPS integral. For any probability level, write the corresponding simulated predictive quantile as shown in the formula, and define the interval score

$$IS_{\alpha}(l,u;y)=(u-l)+\frac{2}{\alpha}(l-y)1\{y<l\}+\frac{2}{\alpha}(y-u)1\{y>u\}$$

The implemented adjusted score for one portfolio-origin-horizon cell is

$$ACRPS_{m,p,o,h} = \frac{0.50\,|y - q_{0.50}| + 0.25\,IS_{0.50}(q_{0.25}, q_{0.75}; y) + 0.10\,IS_{0.20}(q_{0.10}, q_{0.90}; y) + 0.05\,IS_{0.10}(q_{0.05}, q_{0.95}; y)}{0.90}$$

The median absolute-error component has weight 0.50. The central 50%, 80%, and 90% interval scores have weights 0.25, 0.10, and 0.05, respectively, equal to one-half of their exclusion probabilities. The denominator 0.90 is the sum of these four fixed weights. The same quantile levels and weights are applied to every model, portfolio, origin, and horizon; lower values indicate better distributional accuracy.

The final ranking uses the official adjusted CRPS reported in Appendix B. At the origin level, each score compares one predictive distribution with one realized terminal log return. For model-level ranking, origin-level scores are first averaged within each portfolio-horizon pair, and the final score is the mean across those portfolio-horizon pairs. This is the single scoring convention used for all model ranks, figures, tables, and percentage-improvement claims in this paper.

$$Score_m = \frac{1}{|P||H|} \sum_{p \in P} \sum_{h \in H} \left( \frac{1}{|O_{p,h}|} \sum_{o \in O_{p,h}} ACRPS_{m,p,o,h} \right)$$

$$\Delta_m = 1 - \frac{Score_m}{Score_b}$$

This convention prevents short horizons from mechanically dominating the ranking simply because more late-sample origins have observable short-horizon outcomes. It is not a formal test that treats overlapping horizon returns as independent observations. Monte Carlo simulation approximates each model's predictive quantiles, while the realized terminal return and scored cell set are fixed by the historical panel. The empirical ranking is score based. Production selection is made within the leading out-of-sample specifications by requiring competitive adjusted CRPS and favoring explicit integration of stochastic-volatility hyperparameter and terminal-state uncertainty; runtime is not part of the score.

Relative improvement is reported as a skill score against the naive benchmark:

$$Improvement = 1 - \frac{Score_{selected}}{Score_{naive}}$$

where the numerator is the selected model's adjusted CRPS and the denominator is the corresponding naive-benchmark adjusted CRPS. A positive value means the model improves on the naive benchmark; a value of zero means no improvement.

## 6. Model Set And Benchmarks

The documented comparison record contains 211 scored forecasting specifications: the canonical 200-specification model-selection file plus eleven later same-panel state-inference, stochastic-volatility, and multiscale-memory specifications. It is structured over four forecasting ingredients: the expected-return process, the volatility process, the innovation distribution, and the resampling or path-generation rule. Across that grid, specifications include zero-mean and sample-mean baselines, empirical-Bayes and heteroskedasticity-and-autocorrelation-consistent mean estimators, conditional-volatility overlays, latent stochastic-volatility models, deterministic Laplace and integrated nested Laplace approximation state inference, particle Gibbs with ancestor sampling challengers, empirical and parametric innovation laws, tail-refined residual variants, and block-bootstrap simulation rules.

The benchmark used for direct comparison is the naive benchmark: an independent historical portfolio bootstrap. At a forecast origin, let the training set for a portfolio be

$$T_{p,o} = \{ r_{p,t} : t \leq o \}$$

For each simulation path, the benchmark draws daily returns independently with replacement from the origin-specific training set and forms the horizon return

$$\{r_{p,o+j}^{(s)}\}_{j=1}^{h} \sim i.i.d.\, empirical draws from T_{p,o}$$

Thus the naive benchmark uses each portfolio’s own history, but it does not estimate a dynamic mean, condition on volatility, infer latent states, or learn parameters beyond the empirical return distribution available at the origin. This benchmark is intentionally simple. It asks whether a sophisticated probabilistic model adds distributional forecasting value beyond repeatedly sampling from the portfolio’s own realized return history.

Adjusted CRPS provides the common distributional-accuracy measure across the documented specifications. The production specification replaces fixed future multiscale shock mixtures and fixed residual injections with an origin-estimated shared-state law: multiscale component innovations are driven by the central latent-state shock, while residual persistence, residual innovation variance, and residual loading on that shock are estimated from the latent-state history available at the forecast origin. The resulting official adjusted CRPS is 0.229988. Runtime enters only as an implementation constraint after the statistical specification is defined.

### 6.1 Popularized Monte Carlo Methodology Analogs

To make the comparison externally interpretable, the final evidence file also contains representative popularized Monte Carlo methodology analogs. These analogs are grouped by their statistical structure rather than by implementation label. The relevant question is whether the selected model adds distributional forecasting value beyond familiar simulation families when all models are scored on the same portfolio-origin-horizon cells.

The first family is historical resampling. The independent historical bootstrap samples realized portfolio returns with replacement from the training history available at the forecast origin. Its strength is that it preserves the empirical marginal return distribution without imposing a parametric law. Its weakness is that independent draws do not preserve serial dependence, regime persistence, or time-varying volatility.

The second family is parametric simulation. Gaussian simulation uses an estimated mean and variance and then generates normal return shocks; Student-t simulation keeps the same basic structure while allowing heavier tails. These models are transparent and easy to communicate, but their distributional shape is determined primarily by the chosen parametric law.

The third family is block-bootstrap simulation. Rather than drawing each return independently, block-bootstrap methods resample short sequences of empirical returns or standardized residuals. This preserves more local serial dependence and volatility-regime clustering than the independent bootstrap. In the official evidence file, the block-bootstrap analogs include stationary, moving-block, and circular-block variants.

The fourth family is generalized autoregressive conditional heteroskedasticity-based simulation. GARCH(1,1) models update conditional variance as a function of lagged shocks and lagged variance, allowing volatility to cluster through time. The scored analogs include Gaussian and Student-t parametric variants and filtered historical simulation variants that combine conditional volatility with empirical residual resampling.

These analogs are useful benchmarks because they correspond to simple historical sampling, parametric mean-variance simulation, fat-tailed parametric simulation, serial-dependence-preserving bootstrap simulation, and conditional-volatility simulation. They are not given special treatment in ranking; they are ordinary rows in the same final evidence file.

## 7. Selected Forecasting Model

The selected production model is a history-gated calendar-geometric, mode-centered, posterior-weighted integrated nested Laplace quadrature multiscale stochastic-volatility forecast model with a historical annualized-log-growth mean anchor and origin-estimated shared-state residual dynamics. Its role is to generate simulated daily return paths whose cumulative returns define the predictive distribution at each horizon. The model has four separable layers: an anchored latent expected-return process, deterministic latent stochastic-volatility posterior integration, a multiscale predictive summary of the central smoothed volatility path with an estimated residual state, and empirical innovation resampling. Production forecasts require at least 252 clean daily return observations so every forecast uses this documented methodology; when that history is unavailable or the model cannot be fit, no substitute forecasting methodology is used.

### 7.1 Expected-Return Component

The mean component estimates a latent expected-return process with a historical annualized log-growth anchor. At every forecast origin, this anchor is computed only from information available through that origin. For log returns, the origin-time anchor begins with the training-sample daily log mean,

$$\bar{r}_{p,o} = \frac{1}{T_o} \sum_{t \le o} r_{p,t}$$

$$g_{p,o}^{\log,ann} = 252\,\bar{r}_{p,o},\, g_{p,o}^{eff,ann} = \exp\left(g_{p,o}^{\log,ann}\right) - 1$$

The annualized continuously compounded log-growth anchor is 252 times this daily mean. Its corresponding effective annual compound growth rate is obtained by exponentiating the annualized log rate and subtracting one. The implementation expresses returns as a Sharpe-like state by using the training-sample volatility:

$$\sigma_{p,o} = sd\left(r_{p,t} : t \le o\right)$$

$$SR_{anchor} = \mu_{p,o}^{CAGR} \frac{\sqrt{252}}{\sigma_{p,o}}$$

The estimated state is the deviation of the standardized return state from the origin-time anchor. Rather than relying purely on the sample mean, the model treats expected return as uncertain and shrinks it toward an economically interpretable long-run growth estimate. In simulation, the daily drift is scaled by conditional volatility so that expected return behaves like a conditional Sharpe-ratio process rather than a fixed daily mean detached from the volatility state.

Operationally, define the standardized return observation relative to the anchor as:

$$x_t = r_{p,t} \frac{\sqrt{252}}{\sigma_{p,o}} - SR_{anchor}$$

The mean component filters this signal with a first-order autoregressive dynamic linear model:

$$x_t = s_t + \varepsilon_t, \varepsilon_t \sim \mathcal{N}\left(0, R_d\right)$$

$$s_t = \phi_d s_{t-1} + a_t, a_t \sim N\left(0, Q_d\right)$$

where the transition persistence and state-noise ratio are estimated at the forecast origin by Kalman marginal likelihood using the bounded limited-memory Broyden-Fletcher-Goldfarb-

Shanno numerical optimizer over multiple starting points. The filtered terminal state gives an origin-time expected Sharpe deviation. The forecasting procedure then draws a daily mean path from the fitted first-order autoregressive dynamic linear model. Conditional volatility scaling is applied later in the path generator after the multiscale volatility state is simulated:

$$m_o^{\mu} = m_o^{s} \frac{\sigma_{p,o}}{\sqrt{252}}, P_o^{\mu} = P_o^{s} \frac{\sigma_{p,o}^2}{252}, Q_d^{\mu} = Q_d \frac{\sigma_{p,o}^2}{252}$$

$$d_0^{(s)} \sim N\left(m_o^{\mu}, P_o^{\mu}\right)$$

$$d_j^{(s)} = \phi_d d_{j-1}^{(s)} + a_j^{\mu,(s)}, a_j^{\mu,(s)} \sim N\left(0, Q_d^{\mu}\right)$$

$$\tilde{\mu}_j^{(s)} = \mu_{p,o}^{CAGR} + d_j^{(s)}$$

This matters for long horizons. A raw sample mean can dominate simulated 10-year distributions even when the historical estimate is noisy. The selected model reduces that instability by anchoring the return forecast while still allowing state-dependent risk compensation.

### 7.2 Latent Stochastic Volatility

The volatility component models the unobserved log variance of portfolio returns. Daily returns do not reveal conditional variance directly; they provide residual returns whose squared values are noisy log-chi-square measurements of variance. The procedure first forms the residual return relative to the origin-time sample daily log-mean anchor used by the production volatility filter, rescales it into percentage units, applies the theoretical log-chi-square mean correction, and winsorizes the resulting measurement before stochastic-volatility filtering:

$$e_t = 100\left(r_{p,t} - \mu_{p,o}^{CAGR}\right)$$

$$f_o = max\{0.1\, Q_{0.001}(\{e_t^2 : e_t^2 > 0\}), 10^{-10}\}$$

$$y_t^{SV} = clip_{0.005,0.995}\left(\log(max(e_t^2, f_o)) - E[\log \chi_1^2]\right)$$

The factor of 100 is the production implementation's percentage-unit transformation inside the volatility filter; it is not a change to the scored return target. The clipping operator denotes the empirical 0.5% and 99.5% winsorization applied when the origin-time sample is long enough. The realized terminal returns used for adjusted-CRPS scoring are not winsorized by this step.

The latent log-variance process is estimated by the production implementation as a first-order autoregressive state equation with a Gaussian log-chi-square measurement approximation, transformed-parameter posterior-mode optimization, and a Rauch-Tung-Striebel smoothed central latent-state path. The integrated nested Laplace step then constructs posterior-corrected tensor Gauss-Hermite nodes around the local hyperparameter posterior and integrates the conditional Gaussian terminal-state marginal at every node. The state-space equations and transformed posterior objective are:

$$h_t = \ell + \phi(h_{t-1} - \ell) + \eta u_t, u_t \sim N(0,1)$$

$$y_t^{SV} = h_t + v_t, v_t \sim N(0, Var(\log \chi_1^2))$$

$$\theta = (\ell, logit(\phi), \log \eta)$$

$$\log \pi(\theta / y) = \log L_{KF} - \frac{1}{2}\left(\frac{\ell - \overline{y}}{4}\right)^2 - \frac{1}{2}\left(\frac{\phi - 0.94}{0.20}\right)^2 - \frac{1}{2}(\log \eta - \log 0.35)^2 + \log\{\phi(1-\phi)\} + \log \eta + C$$

$$\hat{\theta} = arg\,max_\theta \log \pi(\theta / y), H = -\nabla_\theta^2 \log \pi(\hat{\theta} / y), \Sigma = H^{-1} = L L^\top$$

Here ell is the long-run log-variance level, phi is the first-order autoregressive persistence parameter, and eta is the volatility-state innovation scale. On their natural scales, the implemented prior kernels are Gaussian in ell around the origin-time mean of the corrected log-variance measurements with scale 4, Gaussian in phi around 0.94 with scale 0.20 on the stationary unit interval, and Gaussian in log eta around log(0.35) as a kernel with respect to eta. Optimization and quadrature use the transformed coordinates (ell, logit(phi), log eta), so the transformed log posterior includes the exact Jacobian log{phi(1-phi)} + log eta. In particular, the transformed eta-coordinate contribution is -0.5[log eta - log(0.35)]^2 + log eta, rather than an unadjusted Gaussian density in log eta. Bounded limited-memory quasi-Newton optimization locates the transformed posterior mode. The inverse negative Hessian supplies the local Gaussian scaling matrix. Third-order tensor Gauss-Hermite quadrature creates 27 hyperparameter nodes, and a posterior-ratio correction reweights each node against the transformed Kalman marginal posterior rather than treating the local Gaussian approximation as exact.

$$\mathbf{z}_{\mathbf{i}} = \sqrt{2}\left(x_{i_1}, x_{i_2}, x_{i_3}\right)^{\top}, \theta_{\mathbf{i}} = \hat{\theta} + L\,\mathbf{z}_{\mathbf{i}}, i_d \in \{1,2,3\}$$

$$\tilde{\omega}_{\boldsymbol{i}} = \left(\prod_{d=1}^{3} \frac{w_{i_d}}{\sqrt{\pi}}\right) \exp\left\{\log \pi(\theta_{\boldsymbol{i}} / y) - \log \pi(\hat{\theta} / y) + \frac{1}{2} \boldsymbol{z}_{\boldsymbol{i}}^{\top} \boldsymbol{z}_{\boldsymbol{i}}\right\}$$

$$h_{o,\boldsymbol{i},j} = m_{o,\boldsymbol{i}} + \sqrt{P_{o,\boldsymbol{i}}}\, u_j, \mathbf{u} = \left(-\sqrt{3}, 0, \sqrt{3}\right), \mathbf{v} = \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right)$$

$$\omega_{\boldsymbol{i},j} = \frac{\tilde{\omega}_{\boldsymbol{i}} v_j}{\sum_{\boldsymbol{i}',j'} \tilde{\omega}_{\boldsymbol{i}'} v_{j'}}$$

At each of the 27 hyperparameter nodes, the conditional Gaussian terminal log-variance marginal is represented by three moment-matching nodes with weights 1/6, 2/3, and 1/6, giving 81 posterior-weighted forecast components. The central maximum-a-posteriori parameter fit supplies the historical Rauch-Tung-Striebel smoothed latent path used by the multiscale decomposition. The leverage correlation is estimated once from lagged standardized returns and innovations in that central smoothed path, then held fixed across quadrature components. For each forecast component, the same future state innovations drive its autoregressive log-volatility recursion and the central-mode recursion. In the equation below, a denotes that autoregressive forecast operator. The difference between the component and central-mode recursions is added to the central multiscale forecast. This mode-centered construction reproduces the central multiscale law at the posterior mode while carrying hyperparameter and terminal latent-state uncertainty into forecast simulation. If the finite-difference Hessian cannot be inverted to a valid covariance, its Cholesky factorization fails, or no finite quadrature support remains, the production path deterministically falls back to the corresponding equal-weight Laplace sigma-point support. If valid quadrature points exist but their posterior weights cannot be normalized, those valid points receive equal weight.

### 7.3 Multiscale Volatility Decomposition

After latent volatility is inferred, the selected model computes a history-gated calendar-geometric decomposition from the central maximum-a-posteriori Rauch-Tung-Striebel smoothed log-volatility path. The use of heterogeneous volatility horizons follows evidence that volatility dynamics operate across multiple time resolutions (Müller et al., 1997; Corsi, 2009). The exact scale rule is deterministic and engineered rather than a universal academic standard. Its short half-lives are 5, 21, 63, 126, and 252 trading days; beyond one year, candidate half-lives double successively to 504, 1,008, 2,016, 4,032, 8,064, and so forth. At each forecast origin, the model retains only half-lives no longer than the available latent-state history, so the component count is determined by origin-time information and grows only logarithmically with sample length. The basis deliberately excludes a one-day component. In the notation below, h with superscript med is the central smoothed path, n_h is its length, ell_q is its sample mean, c_t is the centered path, and K(n_h) is the number of admissible scales. Each exponentially smoothed state begins at c_1 and is subsequently mean-centered. Q contains those states, and zeta equals c minus Qb. The implementation fits b with ridge stabilization and then applies signal-strength shrinkage:

$$\ell_q = \frac{1}{n_h} \sum_{t=1}^{n_h} \hat{h}_t^{med}, c_t = \hat{h}_t^{med} - \ell_q$$

$$L_\infty = \{5,21,63,126,252\} \cup \{252 \cdot 2^j : j = 1,2,\ldots\}$$

$$L(n_h) = \{L \in L_\infty : L \leq n_h\}, K(n_h) = |L(n_h)|$$

$$\varphi_k = \exp\{-\log(2)/L_k\}, \alpha_k = 1 - \varphi_k$$

$$m_{k,1} = c_1, m_{k,t} = \alpha_k c_t + (1 - \alpha_k) m_{k,t-1}, q_{k,t} = m_{k,t} - \overline{m}_k$$

$$b = (Q' Q + 0.05\, Var(c) I)^{-1} Q' c$$

$$b_k \leftarrow b_k \frac{s_k}{s_k + median(s_+) + 10^{-8}}, s_k = |b_k| sd(q_k)$$

The fitted latent log-volatility decomposition, component innovations, residual-state dynamics, common-state loading, simulation bounds, and reliability diagnostics are therefore:

$$h_{T+j}^{(s)} = clip_{[-18,18]}\left[h_{T+j}^{MS,(s)} + a_{T+j}(\theta_s, h_{T,s}, \epsilon_{1:j}) - a_{T+j}(\hat{\theta}, \hat{h}_T, \epsilon_{1:j})\right]$$

$$e_{k,t}=q_{k,t}-\phi_k q_{k,t-1}, \omega_k^2=Var(e_{k,t}), \hat{\phi}_\zeta=clip_{(-1,1)}\left(\frac{\sum_t \zeta_{t-1}\zeta_t}{\sum_t \zeta_{t-1}^2}\right)$$

$$u_{\zeta,t}=\zeta_t-\hat{\phi}_\zeta \zeta_{t-1}, \omega_\zeta^2=Var(u_{\zeta,t}), g_t=standardize\left\{\frac{1}{K}\sum_{k=1}^{K}\frac{e_{k,t}}{sd(e_k)}\right\}$$

$$\hat{\gamma}=clip_{[-1,1]}\left\{\frac{\sum_t g_t(u_{\zeta,t}/\omega_\zeta)}{\sum_t g_t^2}\right\}, h_{low}=Q_{0.005}(h^{med})-IQR(h^{med}), h_{high}=Q_{0.995}(h^{med})+IQR(h^{med})$$

During simulation, one leverage-adjusted central state innovation drives every retained multiscale component. The decomposition residual follows a separately estimated first-order autoregressive state. Its innovation combines the central state shock with an independent standard-normal shock using the origin-estimated loading shown above, so the residual innovation retains unit variance without a fixed mixing coefficient. The resulting central multiscale log variance is clipped to its fitted bounds. The mode-centered integrated nested Laplace correction then adds, for each posterior component, the difference between that component's autoregressive log-volatility recursion and the central-mode recursion under the same future state innovations. The corrected log variance is converted to daily return volatility. This layer is a predictive decomposition of inferred latent volatility rather than a separate realized-volatility estimator.

**7.4 Innovation Simulation**

The return-innovation layer uses untruncated, mean-standardized empirical portfolio-return shocks and stationary-bootstrap resampling. The expected block length is determined rather than manually selected: the forecasting procedure applies the fixed-lag Bartlett autocovariance plug-in rule to both the standardized innovation series and its square and uses the larger result. The stationary bootstrap follows the Politis-Romano construction, restarting a block with probability 1/B and otherwise advancing to the next historical innovation index:

$$I_1 \sim Unif\{1,\dots,n\}, I_j = \begin{cases} Unif\{1,\dots,n\}, & \text{with probability } 1/B, \\ 1+(I_{j-1} \bmod n), & \text{with probability } 1-1/B, \end{cases}$$

The empirical innovation pool is constructed from the origin-time portfolio log returns in percentage units after subtracting the sample daily log mean, which is the historical log-growth anchor, and then recentering and rescaling by the pool's sample standard deviation. It is not standardized by fitted conditional volatility, and the pool is not tail-truncated. Serial dependence is retained with the Politis-Romano stationary bootstrap. Its expected block length is selected deterministically: for both the standardized innovation series and its square, the implementation applies the fixed lag cutoff and Bartlett-weighted autocovariance plug-in equations below, then uses the larger of the two resulting lengths. This is the implemented fixed-lag Bartlett plug-in rule.

$$e_t = 100(r_{p,t} - \bar{r}_{p,o}), z_t = \frac{e_t - \bar{e}}{s_e}$$

$$M = min\{max(5, \lceil 2n^{1/3} \rceil), max(2, \lfloor n/4 \rfloor)\}$$

$$\hat{\gamma}_k = n^{-1} \sum_{t=k+1}^{n} (x_t - \bar{x})(x_{t-k} - \bar{x}), a_k = 1 - \frac{k}{M+1}$$

$$\hat{g} = 2\sum_{k=1}^{M} a_k k \hat{\gamma}_k, \hat{d} = \hat{\gamma}_0 + 2\sum_{k=1}^{M} a_k \hat{\gamma}_k$$

$$B_{plug}(x) = max\left\{1, round\left[\left(\frac{2\hat{g}^2}{\hat{d}^2} n\right)^{1/3}\right]\right\}$$

$$B = max\{B_{plug}(z), B_{plug}(z^2)\}$$

Here the resampled z-star sequence is the mean-standardized empirical portfolio-return shock; it is not obtained by dividing returns by fitted conditional volatility. Mu-tilde is the daily mean path drawn from the annualized-log-growth-anchored dynamic linear model, and sigma in the denominator is the training-sample daily volatility used by that mean model. The subsequent equations show the production path generator. Empirical shocks are stationary-bootstrapped; the central volatility innovation is correlated with the lagged return shock through the central-path leverage estimate; all retained multiscale components share that central innovation; and the fitted residual state combines the same innovation with an independent standard-normal remainder through an origin-estimated loading. Simulated daily log returns are clipped to the documented production bounds.

$$\epsilon_j^{(s)} = \hat{\rho}\, z_{j-1}^{*(s)} + \sqrt{1-\hat{\rho}^{,2}}\, v_j^{(s)}$$

$$q_{k,j}^{(s)} = \phi_k\, q_{k,j-1}^{(s)} + \omega_k\, \epsilon_j^{(s)}$$

$$v_{\zeta,j}^{(s)} = \hat{\gamma}\, \epsilon_j^{(s)} + \sqrt{1-\hat{\gamma}^{,2}}\, v_{\zeta,j}^{(s)}, \zeta_j^{(s)} = \hat{\phi}_\zeta\, \zeta_{j-1}^{(s)} + \omega_\zeta\, v_{\zeta,j}^{(s)}$$

$$h_j^{MS,(s)} = clip_{h_{low}}^{h_{high}}\left(\ell_q + \sum_{k=1}^{K(n_h)} b_k\, q_{k,j}^{(s)} + \zeta_j^{(s)}\right)$$

$$\sigma_j^{(s)} = \exp\left(h_j^{(s)}/2\right)/100$$

$$\mu_j^{(s)} = \tilde{\mu}_j^{(s)}\, \sigma_j^{(s)} / \sigma_{p,o}$$

$$r_j^{(s)} = clip\left(\mu_j^{(s)} + \sigma_j^{(s)} z_j^{*(s)}, -1, 1\right)$$

**7.5 Forecast Algorithm**

At each portfolio-origin pair, the selected model can be summarized as the following forecasting algorithm:

1. Estimate the origin-time daily log-mean and corresponding annualized log-growth anchor, sample volatility, and Sharpe-standardized residual signal from the portfolio daily log-return history.

2. Fit the latent Sharpe-deviation dynamic linear model by Kalman marginal likelihood and draw the future daily mean path from the fitted first-order autoregressive state.

3. Construct the percentage-unit log-variance measurement with theoretical log-chi-square mean correction and empirical winsorization.

4. Fit stochastic-volatility state-space parameters by transformed maximum-a-posteriori optimization and recover the smoothed latent log-variance path with the Kalman/Rauch-Tung-Striebel smoother.

5. Construct 27 posterior-corrected third-order Gauss-Hermite nodes over the three transformed stochastic-volatility hyperparameters, combine each with three nodes from its conditional Gaussian terminal-state marginal, normalize the resulting 81 weights, and estimate the lagged-return leverage correlation from the central smoothed state innovations.

6. Construct the history-gated calendar-geometric scale basis from the available latent-state history; fit its latent-volatility decomposition to the central maximum-a-posteriori smoothed path; and estimate the residual persistence, residual

innovation variance, and residual loading on the standardized common component innovation.

7. Draw the untruncated mean-standardized empirical return shocks with the stationary bootstrap, using the larger deterministic Bartlett plug-in block length obtained from the innovation series and its square.

8. Draw one of the 81 stochastic-volatility components for each simulation path according to its normalized posterior weight; drive every retained multiscale state with the leverage-adjusted central state innovation; propagate the origin-estimated residual state with its variance-normalized common and independent innovation; add the posterior component's autoregressive deviation from the central-mode recursion; apply conditional stochastic-volatility Sharpe mean scaling; simulate daily log returns; clip production path returns to [-1, 1]; and aggregate to the 15 target horizons.

The forecast distribution is not generated by one scalar expected return or one scalar volatility. It is produced by the interaction of a data-anchored expected-return state, posterior-weighted integrated nested Laplace stochastic-volatility inference, multiscale volatility decomposition with origin-estimated shared-state residual dynamics, stationary-bootstrap empirical innovations with a deterministic Bartlett plug-in block length, leverage-coupled volatility shocks, and horizon aggregation.

## 8. Empirical Results

The direct production-model comparison gives the following key rows:

| Model description | Status | Adjusted CRPS | Interpretation |
|---|---|---|---|
| Calendar-geometric mode-centered integrated nested Laplace quadrature multiscale stochastic-volatility model with origin-estimated shared-state residual dynamics | Selected | 0.230 | Production specification |
| Naive benchmark | Benchmark | 0.264 | Independent historical resampling |

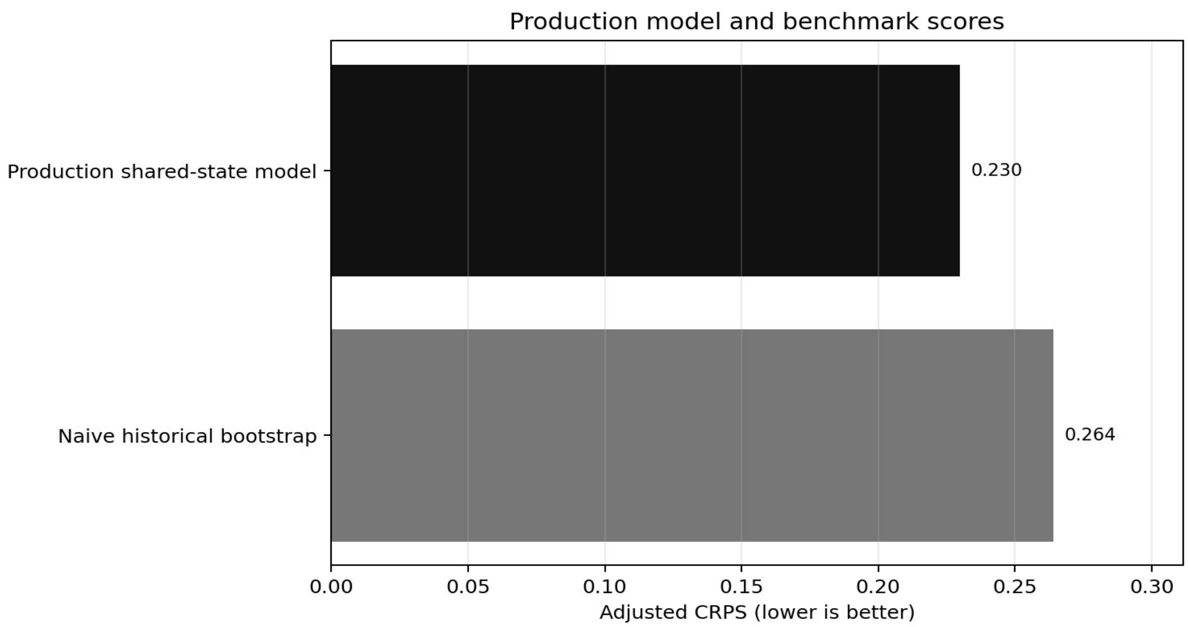


Figure 1. Production-model and naive-benchmark scores. Lower adjusted CRPS indicates better out-of-sample distributional accuracy on realized terminal log returns.

Figure 1 reports the direct benchmark comparison on the 15 fixed-horizon basis. Relative to the naive benchmark, adjusted CRPS falls by 12.90%. This is the headline empirical result for the production specification: it provides better out-of-sample distributional forecasts than simple historical resampling while integrating forecast-relevant stochastic-volatility uncertainty deterministically and estimating its future multiscale shock law from origin-time information.

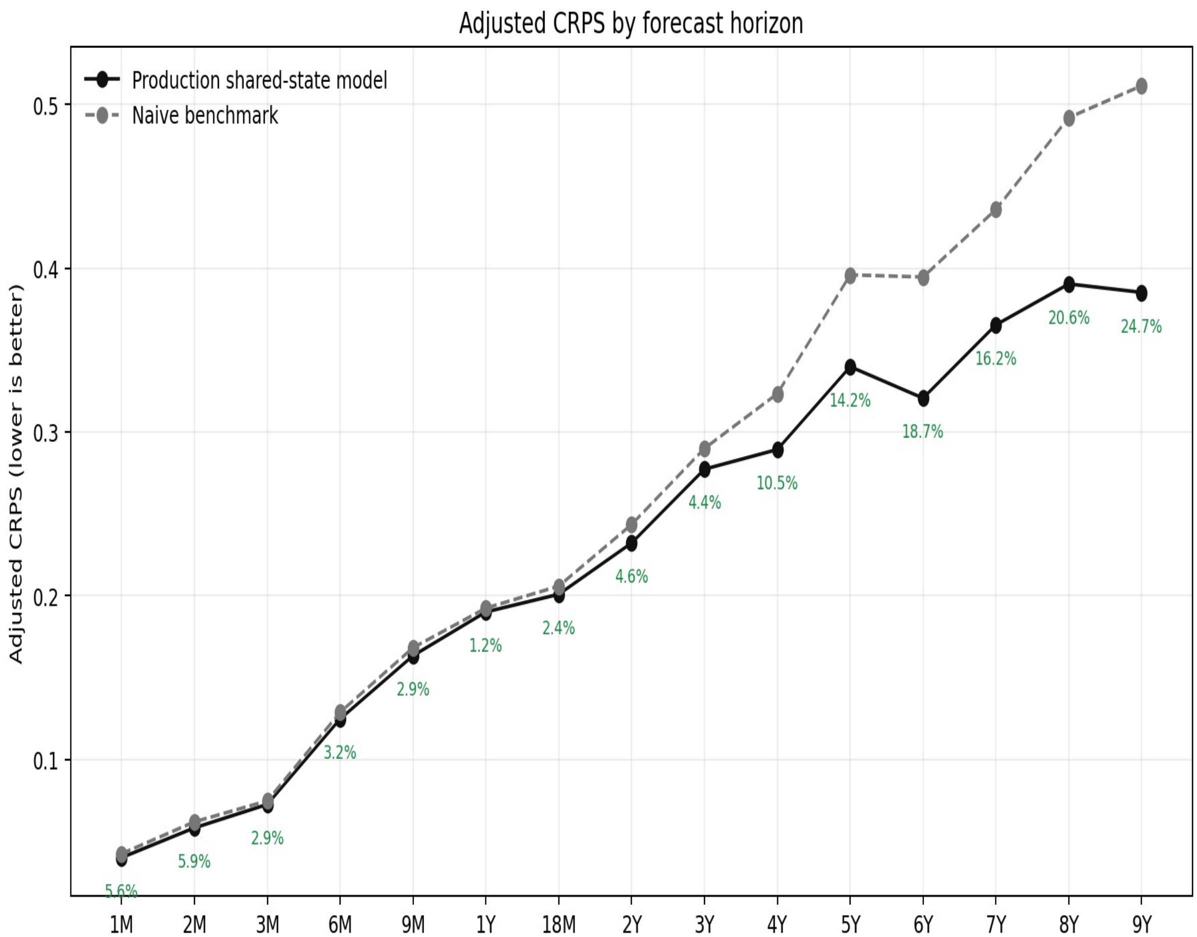


Figure 2. Horizon-by-horizon adjusted CRPS for the production shared-state posterior-weighted integrated nested Laplace quadrature stochastic-volatility model versus the naive historical-bootstrap benchmark. Rows are the fixed horizons from 21 to 2,268 trading days. The green percent labels report production-model adjusted-CRPS outperformance versus the naive benchmark at each horizon.

Figure 2 reports a direct horizon diagnostic. The production model improves on the naive benchmark at every fixed horizon from 1 month through 9 years, with the largest gains concentrated at the longest horizons. The green percentage annotations show the percentage reduction in adjusted CRPS relative to the naive benchmark at each horizon.

The naive benchmark remains competitive within the 211-specification comparison record and outperforms many more elaborate parametric, bootstrap, and conditional-volatility variants. That result is plausible rather than contradictory: independent historical resampling preserves each portfolio's empirical return distribution, avoids unstable long-horizon mean estimation, and does not spend degrees of freedom fitting tail or volatility dynamics that may fail to improve the terminal predictive distribution out of sample. The production model is therefore being compared with a difficult benchmark, not a weak alternative.

Table 2 reports the production forecasting model, structurally distinct leading comparators, and representative popularized Monte Carlo methodology analogs. All rows use adjusted CRPS; lower values are better.

| **Model specification** | **Adjusted CRPS** | Production gain |
|---|---|---|
| Calendar-geometric mode-centered integrated nested Laplace quadrature multiscale stochastic-volatility model with origin-estimated shared-state residual dynamics | 0.230 | - |
| Quasi-maximum-likelihood sigma-point multiscale stochastic-volatility model | 0.231 | 0.48% |
| Omori bivariate-leverage mixture integrated nested Laplace approximation stochastic-volatility model | 0.233 | 1.10% |
| ***Representative benchmark and popularized Monte Carlo analogs*** | | |
| Empirical-Bayes mean with asymmetric GARCH volatility and moving block-bootstrap innovations | 0.252 | 8.87% |
| Constant-mean Student-t baseline | 0.262 | 12.36% |
| Constant-mean Gaussian baseline | 0.263 | 12.42% |

| Model specification | Adjusted CRPS | Production gain |
|---|---|---|
| Naive benchmark (IID historical portfolio bootstrap) | 0.264 | 12.90% |

The production row is followed by structurally distinct stochastic-volatility comparators before the table turns to representative benchmark families and popularized Monte Carlo methodology analogs.

The selected model's advantage is not confined to one fixed-horizon bucket. The table below reports the selected-model versus naive-benchmark comparison by broad horizon group.

| Horizon group | Matched scored cells | Selected model adjusted CRPS | Naive benchmark adjusted CRPS | Improvement vs naive |
|---|---|---|---|---|
| Short horizon (1-3 months) | 7,760 | 0.057 | 0.060 | 4.59% |
| Medium horizon (6-12 months) | 7,440 | 0.159 | 0.163 | 2.31% |
| Long horizon (18 months-5 years) | 8,960 | 0.268 | 0.292 | 8.17% |
| Very long horizon (6-9 years) | 2,160 | 0.365 | 0.458 | 20.29% |
| All horizons | 26,320 | 0.230 | 0.264 | 12.90% |

The horizon-group pattern is consistent with the model’s design. Short horizons leave less room for a state-dependent long-horizon mean and multiscale volatility process to separate itself from historical resampling. The advantage is largest at very long horizons, where drift discipline, volatility-state persistence, and the treatment of volatility regimes compound through the simulated path distribution.

The result is also not confined to one portfolio-construction convention on the fixed-horizon grid. Appendix F splits the comparison by rebalancing rule. The production model improves on the naive benchmark by 17.28% for monthly-rebalance portfolios, 16.56% for quarterly-rebalance portfolios, 18.04% for annual-rebalance portfolios, and 3.93% for buy-and-hold or unrebalanced portfolios. Appendix G reports the comparison separately for each portfolio. At the portfolio level, the production model improves on the naive benchmark in 73 of 80 portfolios, with a median portfolio-level improvement of 14.45%.

## 9. Interpretation

The production model's 12.90% improvement is measured across 26,320 out-of-sample portfolio-origin-horizon forecasts on the fixed 15-horizon grid. The paired summaries average each portfolio's 15 horizon-level loss differences and, as a working assumption, treat the 80 portfolio labels as exchangeable. That assumption is not literal independence: 95.2% of portfolio pairs share at least one constituent, the median pair shares 2 constituents, and horizon outcomes reuse and can overlap the same return histories. The portfolio-label resampling interval for the absolute adjusted-CRPS improvement is [0.0256, 0.0420]. The associated p-values remain conditional sensitivity summaries for the locked panel rather than inference from independent economic units.

**Table 3.** *Paired Out-of-Sample Loss Inference Versus Naive Benchmark*

| Test | Unit | Effect | Statistic | p value | 95% CI |
|---|---|---|---|---|---|
| Paired t-test | Portfolio mean loss difference | 0.0340 | t(79) = 8.06 | p<.001 | [0.0256, 0.0425] |
| Wilcoxon signed-rank | Portfolio paired loss ranks | - | - | p<.001 | - |
| Sign test | Portfolio wins | 73/80 | - | p<.001 | - |
| Portfolio-label bootstrap | 80 portfolio loss differences | 0.0340 | 100,000 draws | p<.001 | [0.0256, 0.0420] |

Note. Positive effects indicate higher adjusted CRPS for the naive benchmark than for the production model. The paired t-test, Wilcoxon signed-rank test, and sign test use 80 portfolio-label loss differences, each averaged across the 15 fixed horizons. The final row is a 100,000-

draw nonparametric bootstrap over those 80 portfolio-label differences. Because portfolio holdings overlap, all four rows are conditional sensitivity summaries for the locked panel rather than inference from independent economic units.

First, the mean process is disciplined. Long-horizon simulations are highly sensitive to drift assumptions. A pure sample mean can overstate confidence in expected return, while a zero-mean model can understate economically persistent compensation. Anchoring the mean to historical annualized log growth while estimating a latent Sharpe-like process is a compromise between those extremes.

Second, volatility is latent and dynamic. Portfolio volatility is not directly observed and is not constant through time. Modeling hidden log variance allows the forecast to condition on volatility clustering without treating noisy squared returns as true volatility.

Third, volatility evolves across multiple time scales. A one-factor volatility process can miss the difference between short-lived volatility shocks and slower volatility regimes. The history-gated decomposition allows weekly, monthly, quarterly, semiannual, annual, and available multiyear volatility components to contribute separately to the forecast path. Its future component states share the central latent-volatility innovation, while an origin-estimated residual state represents variation not explained by the retained basis.

The data-driven diagnostics point to horizon-dependent distributional calibration rather than a single universal source of improvement. Because the naive benchmark already resamples each portfolio's empirical return history, it already carries realized historical skewness, fat tails, and crash observations into the forecast distribution. The selected model's gains are therefore unlikely to be explained by empirical tails alone. The largest improvements occur at the 4- to 9-year fixed horizons, which is more consistent with better joint calibration of the long-horizon drift anchor, latent volatility state, and multiscale volatility persistence. The weaker 1-year improvement is equally informative: it prevents overclaiming that one component, such as the mean model or tail treatment, is solely responsible.

## 10. Limitations And External Validity

The model-selection evidence is strongest as a predictive comparison within the documented evaluation framework. It should not be read as a structural proof about the true data-generating process for portfolio returns.

Several methodological limitations should remain visible. The 211-specification comparison record is broad but not exhaustive, and alternative panels, horizon weighting, scoring rules, or production-selection criteria could change the preferred specification. The popularized Monte Carlo methodology analogs are model-family representatives scored under the same portfolio-origin-horizon target; they should not be interpreted as reproducing every implementation detail used in practice, such as cash-flow timing, inflation indexing, or calendar-year aggregation. The calendar-geometric scale basis is a deterministic multiresolution engineering rule: history determines how many components are admissible, but the calendar anchors and doubling schedule are not uniquely identified economic laws. Finally, shared constituent assets and overlapping horizon returns mean that portfolio-label tests and label-resampling intervals summarize robustness within the locked panel, not uncertainty under independent economic or calendar-time sampling.

The paired significance evidence evaluates the benchmark comparison that matters for production interpretation: the production model versus naive historical bootstrap on the locked panel. Under the documented panel, horizons, origin policy, simulation design, and adjusted-CRPS scoring rule, the production specification attained adjusted CRPS of 0.229988 versus 0.264036 for the naive benchmark.

The supplementary materials provide the following reproducibility records:

1. Appendix A lists holdings, weights, and rebalancing rules for all 80 portfolios.
2. Appendix B lists the 211 model specifications in the documented comparison record and their adjusted CRPS values in a machine-readable comma-separated values file.
3. Appendix C identifies the scoring harness, master files, source artifacts, and replication identifiers.
4. Appendix E reports the selected-model versus naive-benchmark comparison by broad horizon group.
5. Appendix F reports the selected-model versus naive-benchmark comparison by portfolio construction group.
6. Appendix G reports the selected-model versus naive-benchmark comparison separately for each portfolio.

## 11. Conclusion

The rolling-origin evidence supports the history-gated calendar-geometric, mode-centered, posterior-weighted integrated nested Laplace quadrature multiscale stochastic-volatility model with origin-estimated shared-state residual dynamics as the production methodology for long-horizon portfolio simulation. Across 80 portfolios, a common 1979-to-2026 historical window, and the 15 fixed-horizon direct benchmark comparison, the production model reduces adjusted CRPS by 12.90% relative to the naive historical-bootstrap benchmark. Portfolio-label paired analyses support a positive loss advantage within the locked panel, subject to the documented dependence from shared holdings and overlapping return histories. The central conclusion is that careful out-of-sample testing, a common evaluation panel, and structural alignment between the forecast methodology and the long-horizon portfolio problem can identify a model with lower adjusted CRPS than representative historical-bootstrap, parametric Gaussian and Student-t, block-bootstrap, and conditional-volatility simulation analogs under the documented scoring design.

## References


Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics, 31(3), 307-327. https://doi.org/10.1016/0304-4076(86)90063-1

Bracher, J., Ray, E. L., Gneiting, T., & Reich, N. G. (2021). Evaluating epidemic forecasts in an interval format. PLOS Computational Biology, 17(2), e1008618. https://doi.org/10.1371/journal.pcbi.1008618

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics, 7(2), 174-196. https://doi.org/10.1093/jjfinec/nbp001

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica, 50(4), 987-1007. https://doi.org/10.2307/1912773

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance, 48(5), 1779-1801. https://doi.org/10.1111/j.1540-6261.1993.tb05128.x

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359-378. https://doi.org/10.1198/016214506000001437

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. Journal of Business & Economic Statistics, 29(3), 411-422. https://doi.org/10.1198/jbes.2010.08110

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82(1), 35-45. https://doi.org/10.1115/1.3662552

Kim, S., Shephard, N., & Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. Review of Economic Studies, 65(3), 361-393. https://doi.org/10.1111/1467-937X.00050

Martino, S., Aas, K., Lindqvist, O., Neef, L. R., & Rue, H. (2011). Estimating stochastic volatility models using integrated nested Laplace approximations. European Journal of Finance, 17(7), 487-503. https://doi.org/10.1080/1351847X.2010.495475

Müller, U. A., Dacorogna, M. M., Davé, R. D., Olsen, R. B., Pictet, O. V., & von Weizsäcker, J. E. (1997). Volatilities of different time resolutions: Analyzing the dynamics of market components. Journal of Empirical Finance, 4(2-3), 213-239. https://doi.org/10.1016/S0927-5398(97)00007-8

Patton, A., Politis, D. N., & White, H. (2009). Correction to 'Automatic block-length selection for the dependent bootstrap.' Econometric Reviews, 28(4), 372-375. https://doi.org/10.1080/07474930802459016

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. Journal of the American Statistical Association, 89(428), 1303-1313. https://doi.org/10.1080/01621459.1994.10476870

Politis, D. N., & White, H. (2004). Automatic block-length selection for the dependent bootstrap. Econometric Reviews, 23(1), 53-70. https://doi.org/10.1081/ETC-120028836

Rauch, H. E., Tung, F., & Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. AIAA Journal, 3(8), 1445-1450. https://doi.org/10.2514/3.3166

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B, 71(2), 319-392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting, 16(4), 437-450. https://doi.org/10.1016/S0169-2070(00)00065-0

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association, 81(393), 82-86. https://doi.org/10.1080/01621459.1986.10478240

## Supplementary-Material Pointers

Supplementary materials include Appendix A for the portfolio panel, Appendix B for the documented 211-specification adjusted-CRPS score file, Appendix C for the replication trail, and Appendices E-G for horizon, portfolio-construction, and portfolio-level robustness. Appendix data and supporting files are available by request; readers seeking access should contact contact@simfol.io.