

Model Selection for Probabilistic Long-Horizon Portfolio Forecasting

Abstract

This paper evaluates three probabilistic portfolio-forecasting models under a common rolling-origin out-of-sample design. The evidence comprises 80 multi-asset portfolios, 4,080 portfolio-origin tasks, 701,280 eligible portfolio-horizon cells per model, and exact empirical continuous ranked probability score (CRPS) on terminal log returns.

The Frontier model records CRPS 0.256, compared with 0.293 for the Default model and 0.382 for the Competing Public Model. The corresponding reductions are 33.04% relative to the Competing Public Model and 12.67% relative to the Default model. Within the admissible catalog of 175 specifications, the Frontier model ranks first.

1. Introduction

Long-horizon portfolio decisions require predictive distributions rather than point forecasts. The relevant object is the joint law of future asset returns after compounding, rebalancing, and aggregation, because pathwise dependence and marginal uncertainty jointly determine the terminal distribution.

This paper evaluates that object directly. The evaluation applied a common fixed-simulation design at every origin-horizon evaluation. At each forecast origin, every model is estimated exclusively from observations available at the origin, generates an equal number of asset-level paths, applies the recorded portfolio rule, and is scored against subsequently realized terminal log return.

The analysis has three purposes. It compares the two live forecasting models with a deliberately elementary public-method analog; reports the Frontier model's position in a catalog

of 175 admissible specifications; and isolates temporal conditions under which relative performance changes.

2. Related Literature

Rolling-origin evaluation follows the forecasting literature's requirement that model comparison reproduce the information set available at the prediction date (Tashman, 2000). This restriction is material for portfolio simulation because full-sample transformations can transmit future distributional information into earlier origins.

CRPS is a strictly proper scoring rule for an entire predictive distribution (Gneiting and Raftery, 2007). Its empirical form is particularly suitable for Monte Carlo output because it evaluates both location and dispersion without imposing a parametric density on the simulated terminal returns.

The model classes draw on four established lines of research: state-space filtering (Kalman, 1960; Rauch, Tung, and Striebel, 1965), stochastic volatility (Kim, Shephard, and Chib, 1998), low-rank factor representations of cross-sectional dependence, and resampling methods for dependent observations.

The competing benchmark is connected to the stationary bootstrap of Politis and Romano (1994), automatic block-length selection of Politis and White (2004), and the correction of Patton, Politis, and White (2009). It therefore represents a stronger public-method analog than independent return sampling while remaining nonparametric and mechanically transparent.

3. Portfolio Panel

The panel contains 80 multi-asset portfolios observed on a common daily return history from 1979-12-31 through 2026-05-13 (46.37 years). Portfolio membership is determined before model scoring under a common-history requirement.

Every portfolio contains equity, fixed-income, and alternative exposures. The construction rule balances the asset-class representation and varies the constituent combinations and rebalancing conventions, producing cross-sectional heterogeneity without changing the common evaluation window.

Daily simple asset returns are converted to log returns for fitting and path generation, then exponentiated when constituent holdings are advanced. Holdings drift between recorded rebalances and return to target weights after the applicable turnover cost. The identical asset histories, portfolio rules, and origin schedule are applied to all three models.

Table 1. Common temporal diagnostics used for model comparison.

4. Rolling-Origin Forecast Design

For portfolio p , forecast origin o , and horizon h , the realized terminal log return is

$$R_{p,o}(h) = \sum_{k=1}^h \log(1 + r_{p,o+k})$$

At each origin, a model generates $M = 240$ simulated asset paths and maps them through the portfolio rule to terminal log-return draws:

$$X_{p,o}^{(m)}(h) = \log \frac{V_{p,o+h}^{(m)}}{V_{p,o}^{(m)}}, \hat{F}_{p,o,h} = M^{-1} \sum_{m=1}^M \delta_{X_{p,o}^{(m)}(h)}$$

The horizon set contains every daily terminal date supported by the origin and the preserved temporal evaluation rule:

$$H_o = \{h \in \mathbb{N} : o + h \leq T, h \text{ is eligible}\}$$

Diagnostic	Training and evaluation rule
Rolling-origin	Repeated origin-time fits; score each eligible subsequent daily horizon
First-half training	Fit on first half; evaluate all eligible outcomes thereafter
First-quarter training	Fit on first quarter; evaluate all eligible outcomes thereafter
Final-quarter holdout	Fit on first three quarters; evaluate final quarter

The design includes a rolling-origin diagnostic and three fixed temporal partitions. These partitions distinguish repeated real-time-like forecasting from extrapolation after one-half, one-quarter, or three-quarters of the common sample.

Table 2 reports the locked evaluation counts and establishes the denominator of every headline comparison.

Evaluation quantity	Locked value
Portfolios	80
Common history	1979-12-31 to 2026-05-13
Daily return observations	11,687
Portfolio-origin tasks	4,080
Simulations per origin	240
Portfolio-horizon cells per model	701,280

Evaluation quantity	Locked value
Distinct daily horizon aggregates	8,766
Loss	Exact empirical CRPS on terminal log return

Each model is evaluated on 701,280 portfolio-horizon cells spanning 8,766 distinct daily horizons. No model changes the origin schedule, horizon availability, simulation count, or portfolio construction.

The broader selection exercise contains 175 admissible forecasting specifications evaluated under the same scoring definition. The paper uses that catalog only to locate the Frontier model; the empirical exposition concentrates on the three public labels relevant to deployment and external comparison.

All rankings are computed from saved machine-readable result files. Scores are not transcribed into the manuscript source: the build process reads the artifacts, validates the common contract, recomputes every relative reduction, and then populates the tables and figures.

The exercise is pseudo-real-time rather than a claim about contemporaneous data availability for each historical constituent. Its identifying restriction is narrower: no forecast at origin o uses return observations dated after o , and model comparison occurs only on common eligible outcomes.

5. Scoring Rule

For predictive cumulative distribution F and realized terminal log return y , CRPS is

$$CRPS(F, y) = \int_{-\infty}^{\infty} (F(z) - I(y \leq z))^2 dz$$

For Monte Carlo draws $x_1, \dots, x_M$, the exact empirical representation is

$$\hat{CRPS} = \frac{1}{M} \sum_{m=1}^M |x_m - y| - \frac{1}{2M^2} \sum_{m=1}^M \sum_{\ell=1}^M |x_m - x_\ell|$$

The first term measures absolute deviation from the observation. The second subtracts one-half of the average pairwise distance between simulated outcomes, thereby rewarding calibrated dispersion rather than concentration alone. Lower values indicate greater distributional accuracy.

The score is computed on terminal log returns for every eligible portfolio-origin-horizon observation. No quantile approximation, interval weighting, or horizon-specific rescaling enters the reported loss.

Within each portfolio-horizon cell, origin losses are averaged first:

$$\bar{L}_{p,h} = |O_{p,h}|^{-1} \sum_{o \in O_{p,h}} \hat{CRPS}_{p,o,h}$$

Model-level loss then assigns equal weight to the resulting portfolio-horizon cells. This prevents a cell with more eligible origins from receiving greater weight solely because of its calendar location.

Let C denote the set of observed portfolio-horizon cells. The reported model score is

$$\bar{L} = |C|^{-1} \sum_{(p,h) \in C} \bar{L}_{p,h}$$

Temporal partitions are evaluated separately with the same cell-first aggregation. Their scores are diagnostic rather than additive components of the headline result because the partitions contain different eligible origin-horizon structures.

For model A relative to comparator B, proportional CRPS reduction is

$$Reduction(A; B) = 1 - \frac{\bar{L}_A}{\bar{L}_B}$$

A positive reduction indicates that model A has lower out-of-sample loss; a negative value indicates that the comparator is more accurate under the specified diagnostic.

6. Model Set And Benchmarks

The core comparison contains the Frontier model, Default model, and Competing Public Model. Their common evidence comprises 80 portfolios, 4,080 origin tasks, 240 simulations per origin, and identical eligible daily horizons. Live website forecasts use the documented Default model or Frontier model specification selected by the interface; no alternative return model is substituted. The locked OOS simulation count and daily-horizon grid reported here belong to validation, whereas production path counts and requested horizons follow the deployment contract. Each live asset path is cumulated through the requested production horizon before portfolio aggregation.

The Frontier model is an asset-level probabilistic state-space model. It estimates time-varying marginal return states, represents cross-asset dependence through a low-dimensional dynamic Gaussian-score system, propagates filtered state uncertainty, and produces coherent joint asset paths before portfolio aggregation. For the Default model, let $\mathcal{D}_0$ be the complete aligned matrix of pre-origin daily asset log returns.

$$D_o = \{r_t \in \mathbb{R}^N : t \leq o, r_t \text{ complete}\}, \bar{r}_{i,o} = n_o^{-1} \sum_{t \in D_o} r_{i,t}$$

$$s_{i,o}^2 = (n_o - 1)^{-1} \sum_{t \in D_o} (r_{i,t} - \bar{r}_{i,o})^2$$

For the Default model, the pre-origin sample contains only dates on which the complete aligned asset-return vector is finite. For each asset, the sample mean is contracted toward zero as an explicit function of its origin-time t statistic; zero is used when the standard error is not positive.

$$t_{i,o} = \frac{|\bar{r}_{i,o}|}{s_{i,o}/\sqrt{n_o}}, \omega_{i,o} = \frac{t_{i,o}^2}{1 + t_{i,o}^2}, \mu_{i,o} = \omega_{i,o} \bar{r}_{i,o}$$

$$J_{m,k} \stackrel{iid}{\sim} DU\{1, \dots, n_o\}, r_{i,o+k}^{(m)} = r_{i,J_{m,k}} - \bar{r}_{i,o} + \mu_{i,o}$$

One IID row index is shared by every asset, so the Default model retains the observed same-day cross section without fitting a copula or covariance model. The shifted log-return row is exponentiated within the common holdings and rebalancing operator. The model is intentionally simple; its local wins do not establish reliable dominance over the Frontier model.

The Competing Public Model applies the same holdings operator to unshifted historical log-return rows selected by a joint stationary bootstrap. It estimates no mean shrinkage, copula, fitted marginal law, conditional-volatility process, factor system, or tail overlay.

6.1 Popularized Monte Carlo Methodology Analogs

For each asset and each of its centered-return and squared-centered-return series, the Competing Public Model computes the following Politis-White automatic stationary-bootstrap quantities. All sums use the pre-origin sample only.

$$\begin{aligned}\tilde{x}_t &= x_t - \bar{x}, m_x = \min \left[\max \left(5, \lceil 2n^{1/3} \rceil \right), \max \left(2, \lfloor n/4 \rfloor \right) \right] \\ \gamma_x(j) &= n^{-1} \sum_{t=j+1}^n \tilde{x}_t \tilde{x}_{t-j}, w_j = 1 - \frac{j}{m_x + 1} \\ G_x &= 2 \sum_{j=1}^{m_x} w_j j \gamma_x(j), D_x = \gamma_x(0) + 2 \sum_{j=1}^{m_x} w_j \gamma_x(j) \\ b(x) &= \max \left\{ 1, \text{round} \left(\left[\frac{2G_x^2}{D_x^2} n \right]^{1/3} \right) \right\}\end{aligned}$$

The automatic rule returns unit length when the series is too short, has nonpositive variance, or has a numerically unidentified long-run denominator. Otherwise $b(x)$ is given above. The portfolio block length is the maximum across centered returns and squared centered returns, bounded by the available history; the restart probability is its reciprocal.

$$\begin{aligned}L_o &= \min \left[n_o, \max_i \left[b(r_i), b \left((r_i - \bar{r}_{i,o})^2 \right) \right] \right], p_o = L_o^{-1}, I_1 \sim DU \{1, \dots, n_o\} \\ I_k &= \begin{cases} (I_{k-1} \bmod n_o) + 1, & \text{with probability } 1 - p_o, \\ J_k, & \text{with probability } p_o, \end{cases} J_k \sim DU \{1, \dots, n_o\}\end{aligned}$$

The selected row index is shared across assets. A continuation advances one historical row with circular wraparound; a restart draws uniformly from the complete aligned history. The comparator therefore preserves empirical marginal shapes, contemporaneous dependence, and local serial sequences, but contains no conditional state update beyond the sampled block.

7. Selected Forecasting Model

The selected Frontier model treats portfolio forecasting as a joint asset-level latent-state problem. It separates marginal evolution from cross-sectional innovation dependence, estimates both using only the pre-origin record, and simulates daily asset returns before applying the portfolio's path-dependent rebalancing rule.

7.1 Expected-Return Component

For asset i , let r_{it} denote daily log return, and let the origin-time sample mean and standard deviation be $\bar{r}_{i0}$ and $\hat{\sigma}_{i0}$. The expected-return state is estimated in annualized Sharpe units:

$$s_{i,t} = \frac{\sqrt{252} r_{i,t}}{\hat{\sigma}_{i,o}}, \bar{s}_{i,o} = \frac{\sqrt{252} \bar{r}_{i,o}}{\hat{\sigma}_{i,o}}$$

$$y_{i,t} = s_{i,t} - \bar{s}_{i,o} = d_{i,t} + v_{i,t}, v_{i,t} \sim N(0, V_i)$$

The centered Sharpe observation is represented by a scalar latent deviation from the historical log-return anchor. This construction preserves a long-run origin-specific anchor while allowing recent evidence to move the conditional expected return.

$$d_{i,t+1} = \phi_i d_{i,t} + \eta_{i,t+1}$$

$$\eta_{i,t} \sim N(0, W_i), 0 \leq \phi_i < 1$$

Persistence and the stationary signal-to-noise ratio are selected by the origin-time Gaussian marginal likelihood under stability constraints. Conditional on those estimates, inference uses exact scalar Kalman recursions.

Writing a_t and R_t for the one-step state prediction and its variance, the update is

$$a_{i,t} = \phi_i m_{i,t-1}, R_{i,t} = \phi_i^2 P_{i,t-1} + W_i$$

with gain and posterior moments

$$K_{i,t} = \frac{R_{i,t}}{R_{i,t} + V_i}, m_{i,t} = a_{i,t} + K_{i,t} (y_{i,t} - a_{i,t})$$

$$P_{i,t} = (1 - K_{i,t}) R_{i,t}$$

Forecast paths begin from the filtered origin distribution rather than a plug-in terminal state. Each path then advances the stable state equation, so uncertainty in both the terminal state and its subsequent innovations enters the expected-return trajectory.

$$c_{i,o} = \frac{\sigma_{i,o}}{\sqrt{252}}, \delta_{i,o}^{(m)} \sim N(c_{i,o} m_{i,o}, c_{i,o}^2 P_{i,o})$$

$$\delta_{i,o+k}^{(m)} = \phi_i \delta_{i,o+k-1}^{(m)} + \eta_{i,o+k}^{\delta,(m)}, \eta_{i,o+k}^{\delta,(m)} \sim N(0, c_{i,o}^2 W_i)$$

$$\mu_{i,o+k}^{0,(m)} = \bar{r}_{i,o} + \delta_{i,o+k}^{(m)}$$

$$\mu_{i,o+k}^{(m)} = \mu_{i,o+k}^{0,(m)} \frac{\sigma_{i,o+k}^{(m)}}{\sigma_{i,o}}$$

Here $\hat{\sigma}_{i_0}$ is the origin-time sample standard deviation of daily log returns. The last expression preserves the fitted conditional Sharpe signal when the simulated volatility state changes. It is therefore neither a constant drift nor an unconstrained extrapolation of the historical sample mean.

7.2 Latent Stochastic Volatility

Conditional scale is represented by an unobserved log-variance state. The origin-local map $\mathcal{W}_{i_0}$ applies the estimator's finite-sample floor and empirical winsorization to the bias-corrected log-squared observation:

$$x_{i,t} = W_{i,o} \left(\log \left(\left[100 (r_{i,t} - \bar{r}_{i,o}) \right]^2 \right) - E \left[\log X_1^2 \right] \right)$$

$$x_{i,t} = h_{i,t} + \xi_{i,t}$$

$$E(\xi_{i,t})=0, \text{Var}(\xi_{i,t})=\frac{\pi^2}{2}$$

Because the measurement disturbance is centered log-chi-square rather than Gaussian, the state parameters are fitted at the transformed posterior mode. A Gaussian Rauch–Tung–Striebel smoother at that fitted state supplies the log-variance history. The live scale forecast maps that smoothed origin-time history into the multiscale system specified below rather than extrapolating the scalar state equation directly.

The persistent volatility state evolves according to

$$h_{i,t+1}=\alpha_i+\phi_i(h_{i,t}-\alpha_i)+\omega_i u_{i,t+1}$$

$$0\leq\phi_i<1, u_{i,t}\sim N(0,1)$$

$$a_{i,t}^h=\alpha_i+\phi_i(m_{i,t-1}^h-\alpha_i), R_{i,t}^h=\phi_i^2 P_{i,t-1}^h+\omega_i^2$$

$$K_{i,t}^h=\frac{R_{i,t}^h}{R_{i,t}^h+\pi^2/2}, m_{i,t}^h=a_{i,t}^h+K_{i,t}^h(x_{i,t}-a_{i,t}^h), P_{i,t}^h=(1-K_{i,t}^h)R_{i,t}^h$$

$$J_{i,t}^h=\frac{\phi_i P_{i,t}^h}{R_{i,t+1}^h}, h_{i,t}=\hat{m}_{i,t}^h+J_{i,t}^h(\hat{h}_{i,t+1}-a_{i,t+1}^h)$$

Persistence, long-run scale, and leverage dependence are inferred under stability-preserving transformations. Their smoothed origin-time path is the input to the multiscale decomposition; the future scale recursion is therefore the system in Section 7.3 rather than a direct scalar AR(1) continuation.

$$\sigma_{i,o+k}^{(m)}=100^{-1}\exp\left(h_{i,o+k}^{(m)}/2\right)$$

$$r_{i,o+k}^{(m)} = \min \left[1, \max \left[-1, \mu_{i,o+k}^{(m)} + \sigma_{i,o+k}^{(m)} \varepsilon_{i,o+k}^{(m)} \right] \right]$$

$$L_{i,o}^\varepsilon = \max \left\{ b(z_i^{emp}), b(z_i^{emp} \odot z_i^{emp}) \right\}$$

$$Z_{i,mk}^* = Z_{i,I_k^{(m)}}^{emp}, \varepsilon_i = Z_{i,o}(Z_i^*)$$

$$\hat{F}_{i,o,k}(x) = M^{-1} \sum_{m=1}^M I(r_{i,o+k}^{(m)} \leq x)$$

The marginal innovation pool is centered and standardized before resampling. After stationary-bootstrap selection, $\tilde{\mathcal{Z}}_{i0}$ applies the simulation matrix's deterministic recentering, unit-scale normalization, and numerical bounds. The resulting columns form the marginal predictive distributions subsequently rejoined through the cross-asset dependence system.

7.3 Multiscale Volatility Decomposition

Let $\hat{h}_{it}$ denote the smoothed log-variance path for asset i. The model decomposes that path into a long-run level, centered exponentially weighted components, and a residual component:

$$\tilde{h}_{i,t} = \hat{h}_{i,t} - \ell_i, \ell_i = n_o^{-1} \sum_{t \leq o} \hat{h}_{i,t}$$

$$\phi_{i,j} = \exp \left(-\frac{\log(2)}{H_{i,j}} \right), q_{i,j,t} = \phi_{i,j} q_{i,j,t-1} + (1 - \phi_{i,j}) \tilde{h}_{i,t}$$

$$\tilde{b}_{i,o} = (Q_i^T Q_i + \Lambda_{i,o})^{-1} Q_i^T \tilde{h}_i, b_{i,o} = S_{i,o}(\tilde{b}_{i,o})$$

$$e_{i,t} = \tilde{h}_{i,t} - \sum_{j=1}^{K_i} b_{i,j} (q_{i,j,t} - \bar{q}_{i,j})$$

The component persistence coordinates lie on a pre-specified multiscale horizon lattice restricted to scales supported by the origin history. Λ_{i0} is the origin-local diagonal projection penalty and S_{i0} is the deterministic reliability attenuation applied to its coefficient vector; both are functions

only of the smoothed pre-origin path. Each component innovation scale is the square root of its origin-local empirical innovation variance after application of the estimator's numerical variance floor.

$$q_{i,j,o+k}^{(m)} = \phi_{i,j} q_{i,j,o+k-1}^{(m)} + s_{i,j} u_{i,o+k}^{(m)}$$

$$e_{i,o+k}^{(m)} = \rho_i e_{i,o+k-1}^{(m)} + s_{e,i} \left(\lambda_i u_{i,o+k}^{(m)} + \sqrt{1 - \lambda_i^2} v_{i,o+k}^{(m)} \right)$$

$$v_{i,o+k}^{(m)}, \zeta_{i,o+k}^{(m)} \sim N(0,1), v_{i,o+k}^{(m)} \perp \zeta_{i,o+k}^{(m)}$$

For a simulated marginal path, the corresponding multiscale state is projected to the origin-local empirical log-variance interval $[L_{i0}, U_{i0}]$:

$$h_{i,o+k}^{MS,(m)} = \Pi_{[L_{i,o}, U_{i,o}]} \left(\ell_i + \sum_{j=1}^{K_i} b_{i,j} q_{i,j,o+k}^{(m)} + e_{i,o+k}^{(m)} \right)$$

$$\sigma_{i,o+k}^{(m)} = 100^{-1} \exp(h_{i,o+k}^{MS,(m)} / 2)$$

$$u_{i,o+k}^{(m)} = \kappa_i \varepsilon_{i,o+k-1}^{(m)} + \sqrt{1 - \kappa_i^2} \zeta_{i,o+k}^{(m)}$$

$$\varepsilon_{i,o+k-1}^{(m)} \perp \zeta_{i,o+k}^{(m)}, \zeta_{i,o+k}^{(m)} \sim N(0,1)$$

A common state shock propagates across the retained volatility scales, while the residual state receives a correlated common component and an orthogonal disturbance. This construction permits persistent volatility clustering and leverage without treating a single daily scale as sufficient.

7.4 Innovation Simulation

Cross-asset dependence is estimated from aligned historical ranks rather than from raw covariance. Ties receive their average rank, and the fixed interior-rank constant ε_r prevents infinite Gaussian scores. For n_o observations available at origin o , define

$$U_{i,t} = \Pi_{[\varepsilon_r, 1-\varepsilon_r]} \left(\frac{\text{rank}_o(r_{i,t})}{n_o + 1} \right), z_{i,t} = \Phi^{-1}(U_{i,t})$$

The centered Gaussian-score covariance is spectrally decomposed after a fixed positive spectral stabilization $\mathbf{R}_o$. The retained dimension is the smallest admissible dimension satisfying the fixed origin-time variance criterion while preserving an idiosyncratic degree of freedom. Projected factor scores determine diagonal AR transitions and innovation variances; a degenerate score denominator produces a zero transition. The terminal factor distribution is then obtained by the exact Kalman filter.

$$S_z = (n_o - 1)^{-1} Z_c^T Z_c = V \Lambda V^T, \tilde{\Lambda} = R_o(\Lambda), B = V_q \tilde{\Lambda}_q^{1/2}, D = \text{diag}(S_z - B B^T)$$

$$F = Z_c B (B^T B)^{-1}, \phi_j = \Pi_{[-\bar{\phi}, \bar{\phi}]} \left(\frac{\sum_{t=1}^{n_o-1} f_{t,j} f_{t+1,j}}{\sum_{t=1}^{n_o-1} f_{t,j}^2} \right), Q_{jj} = \Pi_{[q_{min}, \infty)} (\text{Var}(f_{t+1,j} - \phi_j f_{t,j}))$$

$$z_t = \bar{z} + B f_t + D^{1/2} \epsilon_t, f_{t+1} = A f_t + Q^{1/2} \eta_{t+1}, \epsilon_t \sim N(0, I_N), \eta_t \sim N(0, I_q)$$

$$a_0 = 0, P_0 = \text{diag} \left(\frac{Q_{jj}}{1 - \phi_j^2} \right), a_t^- = A a_{t-1}, P_t^- = A P_{t-1} A^T + Q$$

$$P_t = \left((P_t^-)^{-1} + B^T D^{-1} B \right)^{-1}$$

$$a_t = P_t \left((P_t^-)^{-1} a_t^- + B^T D^{-1} (z_t - \bar{z}) \right)$$

The terminal factor state is sampled from its filtered Gaussian distribution and propagated forward. Gaussian scores are converted to uniforms, which select linearly interpolated order statistics from each asset's already simulated marginal return column:

$$f_{o+k}^{(m)} = A f_{o+k-1}^{(m)} + Q^{1/2} \eta_{o+k}^{(m)}$$

$$U_{o+k}^{(m)} = \Phi \left(\bar{z} + B f_{o+k}^{(m)} + D^{1/2} \epsilon_{o+k}^{(m)} \right)$$

$$p_{i,k}^{(m)} = (M-1) U_{i,o+k}^{(m)}, j_{i,k}^{(m)} = \lfloor p_{i,k}^{(m)} J \rfloor, j_{i,k}^{+, (m)} = \min(j_{i,k}^{(m)} + 1, M-1), a_{i,k}^{(m)} = p_{i,k}^{(m)} - j_{i,k}^{(m)}$$

$$\tilde{r}_{i,o+k}^{(m)} = (1 - a_{i,k}^{(m)}) r_{i,o+k, (j_{i,k}^{(m)})}^{marg} + a_{i,k}^{(m)} r_{i,o+k, (j_{i,k}^{+, (m)})}^{marg}$$

$$H_{i,0}^{(m)} = w_i^*, V_0^{(m)} = 1$$

$$H_{i,k}^{-, (m)} = H_{i,k-1}^{(m)} \exp \left(\tilde{r}_{i,k}^{(m)} \right), V_k^{-, (m)} = \sum_i H_{i,k}^{-, (m)}$$

$$\tau_k^{(m)} = \sum_i \left| w_i^* - \frac{H_{i,k}^{-, (m)}}{V_k^{-, (m)}} \right|, V_k^{(m)} = \max \left\{ V_k^{-, (m)} \left(1 - \frac{c \tau_k^{(m)}}{2} \right), 0 \right\}, H_{i,k}^{(m)} = w_i^* V_k^{(m)}$$

Here c is the proportional cost per unit of one-way turnover and w^* is the normalized target-weight vector. Outside a scheduled or band-triggered rebalance, $H_{ik} = H_{ik}^-$ and $V_k = V_k^-$. Terminal portfolio log return is $\log(V_{o+h}/V_o)$. This ordering preserves marginal laws, dynamic dependence, weight drift, turnover cost, and rebalancing within the same simulated path.

7.5 Forecast Algorithm

At each portfolio-origin pair, the Frontier model is implemented by the following forecasting sequence:

1. Estimate each asset's origin-time marginal location and scale states from the pre-origin record.
2. Propagate filtered state uncertainty through the stable marginal transition system.
3. Transform aligned historical asset-return ranks to Gaussian scores and estimate their low-dimensional dynamic factor system.
4. Filter and sample the origin-time factor state, then propagate common and idiosyncratic Gaussian-score innovations.
5. Convert simulated Gaussian scores to uniforms and rejoin the separately simulated asset marginal columns by interpolated order statistics.
6. Advance constituent holdings with the rejoined asset log returns.
7. Apply the recorded scheduled or band-triggered rebalancing rule and its turnover cost when active.
8. Repeat the path construction 240 times and retain terminal portfolio log returns for scoring.

The defining distinction is joint conditional evolution. Historical cross-asset ranks inform the dependence law, but future marginal states and the dynamic common-factor state evolve beyond the origin before holdings-level aggregation.

8. Empirical Results

The three-model comparison yields the following model-level scores:

Model	Role	Exact CRPS	Reduction vs competing
Frontier model	Selected	0.256	33.04%

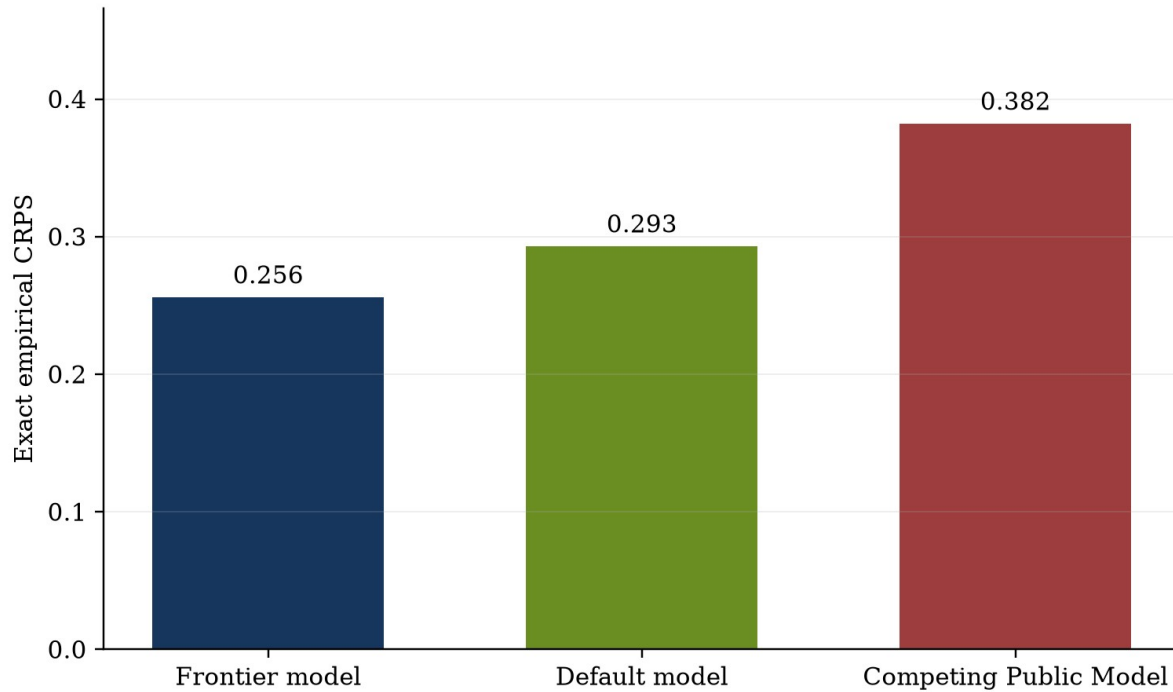


Figure 1. Exact empirical CRPS for the three public models. Lower values indicate greater out-of-sample distributional accuracy.

The Frontier model's CRPS is 0.256, a 33.04% reduction relative to the Competing Public Model (0.382) and a 12.67% reduction relative to the Default model (0.293). The Default model reduces CRPS by 23.33% relative to the competing public-method analog.

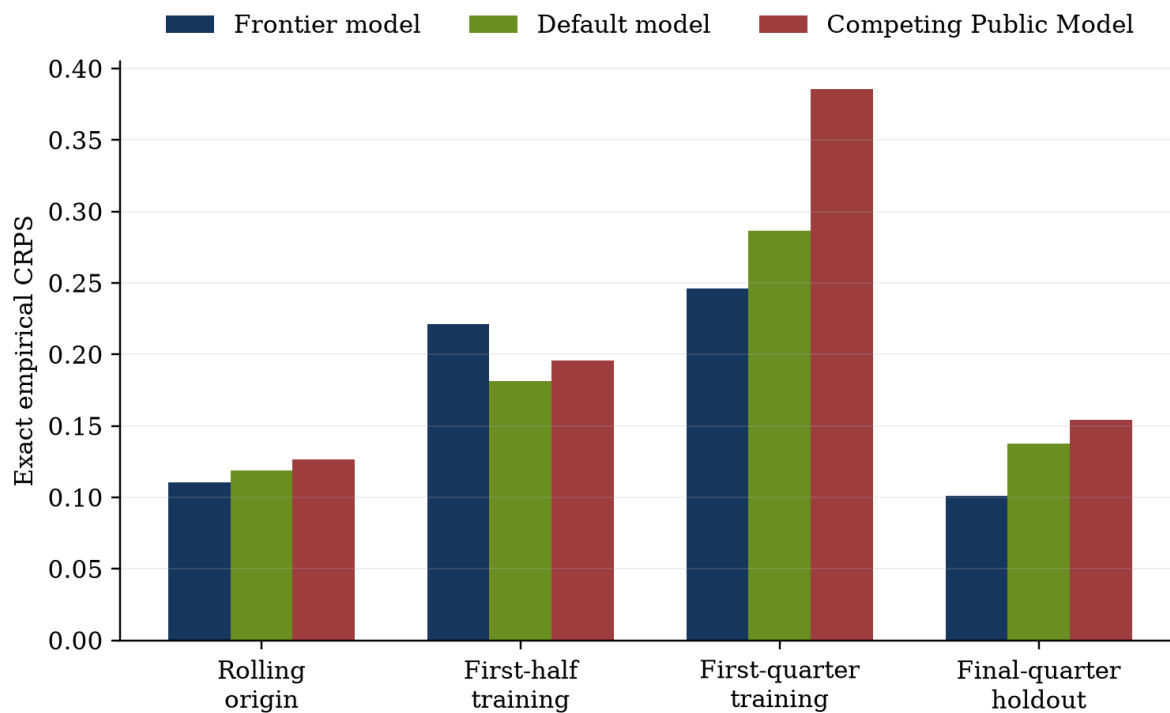


Figure 2. Exact empirical CRPS under four temporal diagnostics. Each bar uses the same model definition and cell-first aggregation; lower values indicate greater accuracy.

The Frontier model is best under the rolling-origin, first-quarter-training, and final-quarter-holdout diagnostics. The Default model is best under first-half training. In that partition the Competing Public Model also records lower loss than the Frontier model, demonstrating that the headline ordering is not mechanically imposed across temporal regimes.

The half-sample reversal does not offset the aggregate result. It instead identifies a regime in which a restrained empirical mean adjustment is particularly effective and the Frontier model's conditional structure is not rewarded by the realized remaining sample.

Table 3 reports the headline scores and proportional reductions. Table 4 identifies the statistical structure of each public model.

Model	Marginal dynamics	Dependence	Simulation object
Frontier model	Filtered location and scale states	Dynamic Gaussian-score factors	Joint asset paths
Default model	Empirical with moderated first moment	Aligned return rows	Joint asset paths
Competing Public Model	Raw empirical returns	Aligned stationary blocks	Joint asset paths

The Frontier model also ranks first among 175 specifications in the canonical out-of-sample catalog. That rank and the three-model comparison use the same exact empirical CRPS definition and equal weighting over portfolio-horizon cells.

Table 5 reports the four temporal diagnostics without collapsing them into the headline score. For the paired predictive-accuracy analysis in Table 6, each model's loss at daily horizon h is first averaged over the 80 portfolio-specific loss cells. The resulting $H = 8,766$ ordered horizon aggregates are paired exactly across models.

Temporal diagnostic	Eligible cells	Frontier	Default	Competing
Rolling-origin	185,360	0.111	0.119	0.127
First-half training	467,520	0.222	0.181	0.196
First-quarter training	701,280	0.246	0.287	0.386
Final-quarter holdout	233,760	0.101	0.137	0.154

$$d_h^{(m)} = \bar{L}_h^{(C)} - \bar{L}_h^{(m)}, \bar{d}_m = H^{-1} \sum_{h=1}^H d_h^{(m)}, H = 8766$$

$$\hat{\gamma}_{m,j} = H^{-1} \sum_{h=j+1}^H \left(d_h^{(m)} - \bar{d}_m \right) \left(d_{h-j}^{(m)} - \bar{d}_m \right)$$

$$\hat{\Omega}_m = \hat{\gamma}_{m,0} + 2 \sum_{j=1}^{L_m-1} \left(1 - \frac{j}{L_m} \right) \hat{\gamma}_{m,j}, z_m = \frac{\sqrt{H} \bar{d}_m}{\sqrt{\hat{\Omega}_m}}$$

9. Interpretation

The principal result is a 33.04% reduction in exact empirical CRPS relative to the Competing Public Model across 701,280 equally weighted portfolio-horizon cells. The paired mean-loss contrast is 0.1262; Bartlett HAC inference gives $z = 6.66$ and one-sided $p = 1.33 \times 10^{-11}$. The dependence-preserving stationary-bootstrap interval is [0.0791, 0.1792].

Table 6. Paired Horizon-Aggregate Predictive-Accuracy Inference

Comparison	Procedure	Effect or statistic	One-sided p	95% interval	Role
Frontier vs competing	Bartlett HAC mean-loss test	Mean $\Delta = 0.1262$; $z = 6.66$	1.33×10^{-11}	[0.0891, 0.1633]	Primary
Frontier vs competing	PW stationary bootstrap	$B = 100,000$; $L = 153$	0.00001	[0.0791, 0.1792]	Robustness
Frontier vs competing	Wilcoxon signed-rank	$W^+ = 38,241,030$	$< 10^{-16}$	—	Breadth
Frontier vs competing	Exact sign test	8,487/8,766 wins	$< 10^{-16}$	—	Breadth
Default vs competing	Bartlett HAC mean-loss test	Mean $\Delta = 0.0891$; $z = 8.13$	2.11×10^{-16}	[0.0676, 0.1106]	Primary
Default vs competing	PW stationary bootstrap	$B = 100,000$; $L = 153$	0.00001	[0.0613, 0.1193]	Robustness
Default vs competing	Wilcoxon signed-rank	$W^+ = 38,335,768$	$< 10^{-16}$	—	Breadth
Default vs competing	Exact sign test	8,436/8,766 wins	$< 10^{-16}$	—	Breadth

Note. At each daily horizon, the loss difference is Competing Public Model CRPS minus comparison-model CRPS; positive values favor the comparison model. HAC uses a Bartlett kernel with bandwidth $L - 1$. $L = 153$ is selected separately for both contrasts by the Politis–White automatic rule. The stationary bootstrap uses geometric blocks, circular continuation, 100,000 replications, seed 20260824, a centered-null one-sided test, and an uncentered percentile interval. Both primary comparisons remain significant after Holm adjustment (adjusted $p < .001$). Wilcoxon and sign results are secondary breadth diagnostics.

First, conditional marginal states contribute information beyond direct historical resampling. The Frontier model allows both expected return and conditional scale to evolve from the origin rather than treating every historical block as exchangeable with the future.

Second, the simulation object is asset-level and joint. Same-day dependence, common shocks, idiosyncratic shocks, weight drift, and scheduled rebalancing interact within each path. This is material when terminal portfolio return is a nonlinear function of the constituent sequence.

Third, complexity alone is not sufficient. The Default model is structurally modest yet materially outperforms the Competing Public Model in the aggregate and wins the half-sample diagnostic. Calibration to the evaluated horizon structure is more important than the number of model components.

The Frontier model nevertheless supplies the more reliable ordering in this evidence: it is best in three of four temporal diagnostics, has the lowest aggregate CRPS, improves on the Competing Public Model at 8,487 of 8,766 daily horizons, and ranks first in the broader admissible catalog. The Default model's local wins should be interpreted as boundary conditions on dominance, not as a general replacement result.

10. Limitations And External Validity

The findings are conditional on the documented panel, common history, origin policy, daily-horizon eligibility rule, simulation count, and CRPS aggregation. Alternative assets, market regimes, transaction assumptions, or decision-specific loss functions may produce a different ordering.

The catalog contains 175 admissible specifications but is not an exhaustive set of probabilistic forecasting models. Rank is empirical within that catalog; it is not a population theorem about all possible estimators.

Monte Carlo error remains because each predictive distribution is represented by a finite simulation sample. The matched simulation count controls this source across models but does not eliminate it. Table 6 does not treat portfolio-horizon cells as independent replications: its unit is the paired cross-portfolio mean at each ordered daily horizon, and its primary procedures preserve local horizon-index dependence. The resulting inference is conditional on the locked horizon grid and the stationary-bootstrap approximation; overlapping outcomes and reused assets preclude a claim of independent economic observations.

The supplementary materials provide the following audit records:

1. Machine-readable source artifacts for each of the three public models.
2. A publication evidence file containing the common contract, computed scores, reductions, temporal diagnostics, paired statistical-validation results, source hashes, and catalog rank.
3. A canonical catalog record defining the admissible comparison set.
4. A retained-format Word manuscript generated from the evidence rather than manually transcribed statistics.
5. A rendered PDF inspected for pagination, equations, tables, figures, and public-label consistency.
6. A deterministic generation program that fails if the OOS contracts or expected manuscript structure diverge.

11. Conclusion

Under the locked rolling-origin design, the Frontier model records exact empirical CRPS 0.256, compared with 0.293 for the Default model and 0.382 for the Competing Public Model. It reduces loss by 33.04% relative to the public-method analog and ranks first among 175

admissible specifications. The result is supported by three of four temporal diagnostics; the half-sample reversal establishes that a well-tuned simple model can prevail locally without providing reliable general dominance.

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](https://doi.org/10.1016/0304-4076(86)90063-1)
- 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>
- Geweke, J., and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies*, 9(2), 557-587. <https://doi.org/10.1093/rfs/9.2.557>
- Gneiting, T., and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359-378. <https://doi.org/10.1198/0162145060000001437>
- Harvey, A. C., Ruiz, E., and Shephard, N. (1994). Multivariate stochastic variance models. *Review of Economic Studies*, 61(2), 247-264. <https://doi.org/10.2307/2297980>
- 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., and 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>

Patton, A., Politis, D. N., and 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., and 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., and 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., and Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8), 1445-1450. <https://doi.org/10.2514/3.3166>

Stock, J. H., and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460), 1167-1179. <https://doi.org/10.1198/016214502388618960>

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](https://doi.org/10.1016/S0169-2070(00)00065-0)

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097-1126. <https://doi.org/10.1111/1468-0262.00152>

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5), 1067-1084. <https://doi.org/10.2307/2171956>

Wright, J. H. (2009). Forecasting US inflation by Bayesian model averaging. *Journal of Forecasting*, 28(2), 131-144. <https://doi.org/10.1002/for.1088>

Zhang, Y., and Wu, W. B. (2021). Gaussian approximation for high dimensional time series. *Annals of Statistics*, 49(4), 1897-1919. <https://doi.org/10.1214/20-AOS2026>

Supplementary-Material Pointers

Supplementary materials comprise the three locked model artifacts, the canonical admissible-catalog record, the generated publication-evidence file, the retained-format Word manuscript, the rendered public PDF, and the deterministic generation program. Public exposition uses only the labels Frontier model, Default model, and Competing Public Model.