AIMS Mathematics, 11(6): 18525–18552.
DOI: 10.3934/math.2026753
Received: 03 April 2026
Revised: 08 June 2026
Accepted: 15 June 2026
Published: 23 June 2026

https://www.aimspress.com/journal/Math

Research article

Monotone functional regression with hybrid depth weighting and block-conformal prediction bands for forward realized variance paths

Çağlar SÖZEN[1], Onur ŞEYRANLIOĞLU[2], Arif ÇİLEK[3] and Abdulmuttalip PİLATİN[4,*]

[1] Görele School of Applied Sciences, Department of Finance and Banking, Giresun University, Giresun, Turkey

[2] Faculty of Economics and Administrative Sciences, Department of Business Administration, Giresun University, Giresun, Turkey

[3] Bulancak School of Applied Sciences, Department of International Trade and Finance, Giresun University, Giresun, Turkey

[4] Department of Finance and Banking, Recep Tayyip Erdoğan University, Rize, Turkey

* **Correspondence:** Email: abdulmuttalip.pilatin@erdogan.edu.tr.

Abstract: We forecast forward realized variance (FRV) paths, defined as cumulative future daily variance proxy curves over a finite trading horizon, using a leakage-disciplined functional framework for multiday risk assessment. The framework combines multiresponse ridge regression, hybrid depth weighting, horizon-weighted blocked cross-validation, and isotonic post-projection to preserve the monotone structure of FRV paths. Uncertainty is summarized through upper one-sided block-calibrated conformal bands, interpreted as empirical risk envelopes under temporal dependence rather than exact distribution-free guarantees. In a fixed panel design for four liquid exchange-traded funds, GDX, GDXJ, XLE, and UUP, over the period 2010–2025, the proposed model reduces long-horizon mean squared error relative to rolling historical FRV by approximately 31.8%, 20.4%, 36.5%, and 28.0%, respectively, over $h = 20{:}30$. Comparisons with heterogeneous autoregressive (HAR) ridge and functional principal component autoregressive (FPCA-AR) benchmarks are asset-dependent. The proposed model is most favorable for GDX and remains close to HAR ridge for GDXJ, whereas HAR ridge and FPCA-AR remain competitive for XLE and UUP. Coverage is conservative or close to nominal at $\alpha = 0.05$ but more heterogeneous at $\alpha = 0.10$. Robustness checks support a cautious interpretation of the method as a shape-aware enhancement of rolling FRV forecasting.

Keywords: forward realized variance; functional regression; depth weighting; isotonic projection; block-conformal prediction; liquid exchange-traded funds
Mathematics Subject Classification: 62M10, 62M20, 62P20, 91G70

1. Introduction

Forward realized variance (FRV) paths summarize the horizon-by-horizon accumulation of future variance risk over a multiday window. For a forecast origin t, the FRV path records cumulative future variance proxies $y_{t,h}$, $h = 1, \ldots, H$ and therefore aligns naturally with operational risk windows such as weekly and monthly margining, stress monitoring, and cumulative loss-control horizons. This perspective is connected to the realized volatility literature, which emphasizes the persistence, aggregation, and forecastability of variance dynamics across horizons [1–4]. Unlike a scalar one-step-ahead volatility forecast, an FRV path describes how risk accumulates across the forecast horizon. This pathwise object is also structurally constrained: Because it is cumulative, it is nondecreasing in h. A forecasting method for FRV should therefore respect both information timing and the monotone geometry of the target.

This paper develops a leakage-disciplined and shape-aware forecasting framework for daily FRV proxy paths. The specification combines four components. First, multiresponse ridge regression (MRR) provides a pathwise shrinkage estimator which is consistent with the use of regularization in high-dimensional and multiresponse forecasting problems [5]. Second, hybrid depth weighting (HDW) combines shape information (based on the geometry of normalized FRV increments) with scale information (based on terminal cumulative variance) to reduce the leverage of atypical training episodes. This component is motivated by robust functional-depth ideas, where curve centrality and outlyingness are used to stabilize inferences for functional observations [6–8]. Third, horizon-weighted blocked cross-validation (HW-BCV) emphasizes long-horizon cumulative risk while preserving chronological validation, in line with time-series validation principles that avoid random-fold look-ahead [9, 10]. Fourth, isotonic post-projection (IPP) enforces the monotonicity of the fitted FRV path through projection onto the monotone cone [11–13]. Uncertainty is summarized with upper one-sided block-calibrated conformal bands (BCC bands), interpreted empirically under temporal dependence rather than as exact distribution-free guarantees under arbitrary dependence [14–16]. These short forms are used in the remainder of the paper.

The empirical design uses a prespecified fixed panel and a chronological evaluation protocol. The panel consists of four liquid exchange-traded funds (ETFs): GDX, GDXJ, XLE, and UUP. The sample spans the years 2010–2025, providing a long daily evaluation window that includes several market conditions. Forecast origins are split chronologically into training, calibration, and test blocks. Model fitting, scaling, depth construction, regime-threshold estimation, and tuning use training information only; the calibration block is reserved for conformal quantile estimation; and the test block is used only for final evaluation. The main horizon is $H = 30$, corresponding to an approximately monthly trading risk window, and $H = 10$ and $H = 20$ are reported as robustness checks.

The target is constructed from daily variance proxies. The main specification uses close-to-close squared returns because they are transparent and widely available, but they are necessarily noisier than high-frequency realized measures. The paper therefore does not claim that daily squared returns provide measurement-equivalent realized variance relative to intraday estimators. Instead, the object is a daily observable FRV proxy path. To examine sensitivity to this measurement choice, the analysis is repeated with daily open–high–low–close (OHLC)-based proxies, including Parkinson, Garman–Klass, and Rogers–Satchell variance measures [17–19]; the corresponding proxy robustness evidence is summarized in Figure A1. Accordingly, the contribution is a forecasting and calibration framework

for daily FRV proxy paths, not a replacement for high-frequency realized volatility measurement.

The empirical findings are interpreted cautiously. The most consistent result is that the proposed depth-weighted model has lower long-horizon loss than rolling historical FRV across all four assets in the long-horizon window $h = 20{:}30$. The corresponding percent reductions in the long-horizon mean squared error (MSE) are approximately 31.8% for GDX, 20.4% for GDXJ, 36.5% for XLE, and 28.0% for UUP. These results indicate lower long-horizon loss than a simple historical curve benchmark. However, the evidence is more heterogeneous relative to heterogeneous autoregressive (HAR) ridge and functional principal component autoregressive (FPCA-AR) benchmarks. The proposed model is most favorable for GDX and remains close to HAR ridge for GDXJ, while HAR ridge and FPCA-AR remain viable alternatives for XLE and UUP. The interpretation is therefore asset-dependent rather than uniformly favorable to one model.

The conformal layer is also interpreted with caution. Standard split-conformal validity relies on exchangeability, which is not literally satisfied in serially dependent financial data. We therefore present the one-sided bands as empirical block-calibrated risk envelopes. Coverage is generally conservative or close to nominal at the stricter $\alpha = 0.05$ level, whereas $\alpha = 0.10$ reveals more asset-specific behavior, with UUP being the most challenging case. Guard factor and block size sensitivity, including the unguarded case $\gamma = 1.00$, are reported to document the coverage–width trade-off rather than to assert a new finite-sample validity theorem.

The paper contributes to the literature in four ways. First, it treats FRV as a monotone curve-valued forecasting target and combines MRR with IPP to preserve the cumulative structure of the path, connecting cumulative risk forecasting with functional data analysis and function-on-scalar/multiresponse regression ideas [20–22]. Second, it introduces HDW to use both shape and scale information when moderating the influence of atypical variance episodes. Third, it uses HW-BCV to align model selection with long-horizon cumulative risk while respecting time order. Fourth, it evaluates uncertainty through BCC bands and reports sensitivity analyses under temporal dependence. These components form a compact forecasting pipeline that remains interpretable while making explicit the information timing, benchmark scope, horizon choice, proxy dependence, and conformal calibration.

The remainder of the paper is organized as follows. Section 2 reviews related work on functional data analysis, realized volatility forecasting, robust depth methods, and conformal prediction under dependence. Section 3 presents the forecasting and calibration methodology. Section 4 describes the data, design, benchmarks, figures, and empirical results. Section 5 concludes with limitations and future research directions.

2. Related works

Functional data analysis (FDA) provides tools for representing, smoothing, and regressing curve-valued observations [20,21,23,24]. In the present setting, FRV paths are observed on a discrete horizon grid and can be treated as multiresponse realizations of an underlying curve, consistent with function-on-scalar regression practice. A defining feature of FRV is monotonicity in the horizon h, because the path is a cumulative sum of future variance proxies. Shape constraints such as monotonicity can be enforced during estimation or through projection. Isotonic regression and the pool-adjacent-violators algorithm provide an efficient projection onto the monotone cone [11–13]. For cumulative variance

targets, such projections act as a structural discipline that improves interpretability without imposing a fully parametric dynamic model.

Robustness to atypical episodes in functional settings is often studied through notions of depth, which quantify centrality and identify outlying curves in function space [6, 7, 24]. Band depth and related measures emphasize geometric centrality, while shape-outlyingness approaches distinguish between amplitude and shape deviations [8]. The present paper uses these ideas to construct a simple hybrid depth-weighting rule for FRV paths. The shape channel measures deviations in normalized increment geometry, whereas the scale channel measures deviations in terminal cumulative variance. The rule is intentionally parsimonious. It is not presented as a uniformly better depth, and Figure A2 and Table A1 therefore report shape-only, scale-only, and hybrid sensitivity results.

Realized volatility forecasting is a central area in empirical finance. High-frequency realized measures provide more efficient estimates of integrated variance than daily squared returns, and the realized volatility literature has documented strong persistence and multiscale dependence in volatility dynamics [1–3]. The heterogeneous autoregressive model of realized volatility (HAR-RV) captures daily, weekly, and monthly components in a simple and effective way [4]. The present paper differs from scalar volatility forecasting by focusing on cumulative forward variance paths. The HAR ridge benchmark retains the multiscale logic of HAR, while the proposed method operates on a curve-valued cumulative target.

The use of daily variance proxies requires careful interpretation. Close-to-close squared returns are transparent and broadly available but noisy. Range-based daily estimators, including Parkinson, Garman–Klass, and Rogers–Satchell measures, use additional OHLC information and can provide alternative daily variance proxies [17–19]. Rather than treating any single daily proxy as definitive, the empirical design reports proxy robustness. This positioning is cautious: The paper studies forecasting of daily observable FRV proxy paths and leaves the integration of high-frequency realized measures to future work.

Conformal prediction provides finite-sample marginal coverage under exchangeability [14, 15]. In financial time series, exchangeability is generally not a credible assumption, and recent work has therefore considered adaptive, weighted, or distribution-shift-aware conformal strategies [16, 25]. The present paper uses a block-max calibration rule to produce a single upper threshold for the whole FRV path. The resulting bands are reported as empirical block-calibrated risk envelopes under temporal dependence. Moving-block bootstrap summaries are also used to describe uncertainty in efficiency gains under serial dependence [26, 27].

Time-series model selection must respect chronology. Random folds can leak future information into the training process and produce overly optimistic forecast assessments. Blocked or rolling validation schemes are therefore preferred for dependent data [9, 10]. The blocked cross-validation used here follows this principle and adds a horizon-weighted loss so that model selection reflects the operational importance of the long-horizon part of the cumulative risk path. This connects functional forecasting practice with the multiscale perspective of volatility modeling while keeping the evaluation leakage-disciplined.

Taken together, the paper combines shape-constrained functional regression, hybrid depth reweighting, horizon-aware blocked tuning, and block-calibrated one-sided uncertainty summaries for FRV path forecasting. The combination is designed to remain interpretable, but the empirical claims are intentionally modest: The proposed method has lower loss than rolling historical FRV, whereas

the magnitude and direction of gains relative to HAR ridge and FPCA-AR depend on the asset and regime.

3. Method

3.1. Overview, terminology, and functional viewpoint

Figure 1 summarizes the forecasting workflow. The object of interest is a forward realized variance (FRV) path, defined as a cumulative future variance proxy curve over a finite trading horizon. Let $v_t \geq 0$ denote a daily variance proxy available at the end of day t. The FRV target is given by Eq (3.1):

$$y_{t,h} = \sum_{j=1}^{h} v_{t+j}, \qquad h = 1, \ldots, H. \tag{3.1}$$

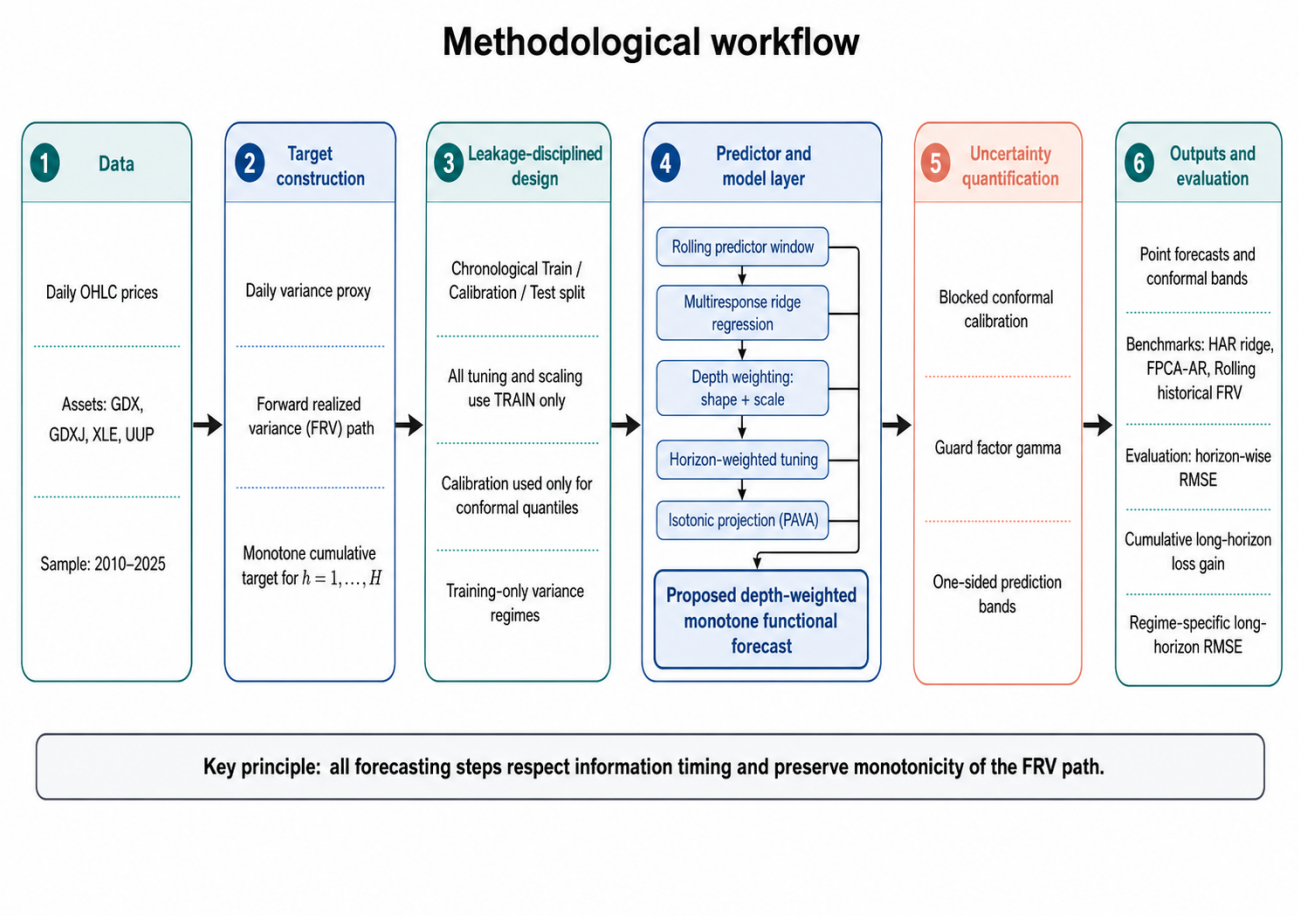

Figure 1. Methodological workflow of the FRV forecasting framework. The figure summarizes target construction, training-only model selection and regime construction, calibration-only conformal thresholding, and final test evaluation.

The main empirical design uses $H = 30$, corresponding to an approximately monthly trading risk window, and the robustness analysis repeats the experiment for $H \in \{10, 20, 30\}$. Because $y_{t,h}$ is a

cumulative sum, the FRV path is nondecreasing in h by construction. Throughout the paper, we therefore use the term variance consistently. If a volatility scale path is desired, it can be obtained by the deterministic transformation $\sqrt{y_{t,h}}$, but estimation, tuning, calibration, and evaluation are conducted on the cumulative variance scale.

The notation v_t is used deliberately because the empirical design considers several daily variance proxy constructions. The main specification uses close-to-close squared returns, $v_t = r_t^2$, which are simple and widely available but noisy relative to high-frequency realized measures. Accordingly, the paper does not claim that daily squared-return FRV paths are measurement-equivalent to intraday realized variance. Rather, the target is interpreted as a daily observable variance proxy path. To examine the sensitivity of the conclusions to this measurement choice, the analysis is repeated with OHLC-based daily proxies, including Parkinson, Garman–Klass, and Rogers–Satchell range-based variance measures [17–19]; the corresponding proxy robustness evidence is summarized in Figure A1. The contribution is therefore a leakage-disciplined functional forecasting design for daily FRV proxy paths, with measurement uncertainty handled through proxy robustness checks rather than through a claim of high-frequency realized volatility equivalence [1–3].

The estimator is defined on the discrete grid $h \in \{1, \ldots, H\}$, so the problem can be treated as finite-dimensional multiresponse regression. This is consistent with function-on-scalar regression practice, where a latent curve $Y_t(s)$ is observed on a grid $\{s_h\}$ and modeled through its evaluations $\mathbf{y}_t = (y_{t,1}, \ldots, y_{t,H})^\top$ [20, 22, 28]. The isotonic post-projection used below is the discrete analog of projecting an underlying monotone function onto a monotone cone in L_2, and it is natural here because FRV is intrinsically monotone across the forecast horizon.

The framework has four main components. First, a multiresponse ridge specification provides a simple shrinkage baseline for the whole FRV path. Second, a hybrid depth-weighting rule combines shape and scale information to moderate the influence of atypical training episodes. Third, horizon-weighted blocked cross-validation emphasizes cumulative risk at longer horizons. Fourth, upper one-sided block-calibrated conformal bands (BCC bands) summarize uncertainty at the path level. The empirical analysis compares the proposed specification with rolling historical FRV, HAR ridge, FPCA-AR, and generalized autoregressive conditional heteroskedasticity (GARCH)-type benchmarks, and it reports ablation, horizon sensitivity, proxy robustness, regime-specific, Diebold–Mariano, and model confidence set (MCS)-style robustness evidence.

3.2. Monotone multiresponse ridge baseline and isotonic post-projection

Let $\mathbf{x}_t \in \mathbb{R}^p$ denote the predictor vector available at the forecast origin. In the main specification, $\mathbf{x}_t$ includes HAR-style summaries of past daily variance proxies and lagged FRV-related summaries. Stack responses and regressors over the training index set $\mathcal{I}_{\text{train}}$:

$$Y = [\mathbf{y}_t]_{t \in \mathcal{I}_{\text{train}}} \in \mathbb{R}^{n \times H}, \qquad Z = [\mathbf{x}_t^\top]_{t \in \mathcal{I}_{\text{train}}} \in \mathbb{R}^{n \times p}.$$

An intercept is included and is not penalized. Let $P = \operatorname{diag}(0, 1, \ldots, 1) \in \mathbb{R}^{p \times p}$. The unweighted multiresponse ridge estimator solves Eq (3.2):

$$\widehat{B}_\lambda = \arg\min_{B \in \mathbb{R}^{p \times H}} \|Y - ZB\|_F^2 + \lambda \operatorname{tr}(B^\top P B) = (Z^\top Z + \lambda P)^{-1} Z^\top Y. \tag{3.2}$$

The corresponding fitted paths are

$$\widehat{\mathbf{y}}_t = \mathbf{x}_t^\top \widehat{B}_\lambda \in \mathbb{R}^H.$$

This common-λ multiresponse specification is used as a transparent pathwise shrinkage estimator [5]. Because the variance scale increases with the horizon, Table A2 also reports a horizon-specific regularization variant. Thus, the empirical assessment does not rely only on an unexamined common-penalty restriction.

Because FRV targets are cumulative sums, the true path is monotone in h. We enforce this shape constraint by projecting each fitted path onto the monotone cone

$$C = \{\mathbf{u} \in \mathbb{R}^H : u_1 \le \cdots \le u_H\}.$$

The projected forecast is obtained from Eq (3.3):

$$\widehat{\mathbf{y}}_t^{\mathrm{iso}} = \Pi_C(\widehat{\mathbf{y}}_t) = \arg\min_{\mathbf{u}\in C} \|\mathbf{u} - \widehat{\mathbf{y}}_t\|_2^2. \tag{3.3}$$

This projection is computed by the pool-adjacent-violators algorithm [11–13]. We write $\widehat{y}_{t,h}^{\mathrm{iso}}$ for the hth component of the projected path. The projection is a shape-preserving post-processing step and uses neither calibration nor test information.

3.3. Hybrid depth weights: Shape and scale channels

Financial variance paths contain bursts, stress episodes, and scale shifts. To reduce the leverage of atypical training episodes while retaining a simple estimator, we assign observation-level weights using a hybrid depth score. The score combines a shape channel (based on the geometry of normalized FRV increments) and a scale channel (based on the terminal FRV level). This construction is motivated by robust functional depth and shape-outlyingness ideas [6–8, 24]. We do not claim that this particular depth rule is uniformly preferable to established functional depths. Instead, its empirical role is examined through ablation and depth channel sensitivity checks.

For robust scaling, we use the median absolute deviation (MAD) and its normal-consistency-scaled version (MADN). For a univariate sample $\{u_i\}$, let

$$\mathrm{MAD}(u) = \mathrm{median}\{|u_i - \mathrm{median}(u)|\}, \qquad \mathrm{MADN}(u) = 1.4826\,\mathrm{MAD}(u).$$

If $\mathrm{MADN}(u)$ is numerically degenerate, the sample standard deviation is used as a fallback, and then one if needed. This convention prevents unstable standardization when a short segment contains nearly identical observations.

Let

$$\Delta\mathbf{y}_t = (y_{t,1},\, y_{t,2} - y_{t,1},\, \ldots,\, y_{t,H} - y_{t,H-1})$$

denote the forward increment vector. The increments are normalized into a probability vector,

$$p_{t,h} = \frac{\Delta y_{t,h}}{\sum_{j=1}^{H} \Delta y_{t,j}},$$

with a uniform fallback if the denominator is numerically degenerate. Let $c_{t,h} = \sum_{j=1}^{h} p_{t,j}$. On a fixed grid $\mathcal{U} = \{u_\ell\}_{\ell=1}^m \subset (0,1)$, define the discrete quantile function $Q_t(u)$ by linear interpolation of the inverse of $c_{t,h}$. Let $Q_{\mathrm{med}}(u)$ be the pointwise median of $\{Q_t(u)\}$ over the training set. The shape distance is

$$d_t^{\mathrm{shp}} = \left[\frac{1}{m}\sum_{\ell=1}^{m}\{Q_t(u_\ell) - Q_{\mathrm{med}}(u_\ell)\}^2\right]^{1/2}. \tag{3.4}$$

Distances are converted to bounded shape scores using training-only robust scaling and high-quantile clipping:

$$D_t^{\mathrm{shp}} = \exp\left\{-\left[\frac{\min\{d_t^{\mathrm{shp}}, q_{\mathrm{clip}}^{\mathrm{shp}}\}}{s_{\mathrm{shp}}}\right]^2\right\}, \qquad s_{\mathrm{shp}} = \mathrm{MADN}\{d_t^{\mathrm{shp}} : t \in \mathcal{I}_{\mathrm{train}}\},$$

where $q_{\mathrm{clip}}^{\mathrm{shp}}$ is the empirical training quantile used for clipping.

Let $s_t = \log(1 + y_{t,H})$ denote the log terminal FRV level. Define

$$z_t = \frac{|s_t - \mathrm{median}(s)|}{\mathrm{MADN}(s)}, \qquad s = \{s_t : t \in \mathcal{I}_{\mathrm{train}}\}.$$

After clipping z_t at a high training quantile $q_{\mathrm{clip}}^{\mathrm{scl}}$, the scale score is

$$D_t^{\mathrm{scl}} = \exp\left[-\min\{z_t, q_{\mathrm{clip}}^{\mathrm{scl}}\}^2\right].$$

The two channels are combined multiplicatively:

$$D_t = (D_t^{\mathrm{shp}})^a (D_t^{\mathrm{scl}})^{1-a}, \qquad a \in [0, 1]. \tag{3.5}$$

The resulting scores are mildly bounded and normalized to have mean one on the training set:

$$\frac{1}{|\mathcal{I}_{\mathrm{train}}|} \sum_{t \in \mathcal{I}_{\mathrm{train}}} w_t = 1.$$

The lower and upper bounds prevent a small number of observations from receiving nearly zero or excessively large leverage. The tuning grid for a, the clipping rule, and the selected values are reported in Table A3.

Let $W = \mathrm{diag}(w_t)$ and keep the same intercept/nonpenalization matrix P. The depth-weighted multiresponse ridge estimator solves Eq (3.6):

$$\widehat{B}_{\lambda,W} = \arg\min_B \|W^{1/2}(Y - ZB)\|_F^2 + \lambda\,\mathrm{tr}(B^\top P B) = (Z^\top W Z + \lambda P)^{-1} Z^\top W Y. \tag{3.6}$$

Predictions are

$$\widehat{\mathbf{y}}_t = \mathbf{x}_t^\top \widehat{B}_{\lambda,W},$$

followed by the same isotonic projection $\widehat{\mathbf{y}}_t^{\mathrm{iso}}$. The depth weights are estimated using training information only.

3.4. Horizon-weighted blocked cross-validation

Hyperparameters are selected using time-respecting blocked cross-validation within the training block. The training sample is partitioned into contiguous folds. For a given validation fold, model fitting, scaling, and depth-weight computation use only the corresponding fold-training observations. The validation fold is then used only to evaluate candidate hyperparameters. This avoids random-fold look-ahead and preserves the chronological structure of the forecasting problem.

To emphasize cumulative risk at longer horizons, candidates are evaluated with a horizon-weighted validation loss,

$$\mathcal{L} = \frac{1}{|\mathcal{V}|} \sum_{t \in \mathcal{V}} \sum_{h=1}^{H} \omega_h (y_{t,h} - \widehat{y}_{t,h}^{\mathrm{iso}})^2. \tag{3.7}$$

Here, $\mathcal{V}$ denotes the validation fold and

$$\omega_h \propto h^{\beta}\{1 + \eta\, \mathbf{1}(h \geq h_0)\}, \qquad \sum_{h=1}^{H} \omega_h = 1.$$

For the main $H = 30$ specification, $h_0 = 20$, corresponding to the long-horizon portion of the monthly FRV path. The parameters β and η govern the strength of long-horizon emphasis.

For each candidate (a, β, η), the ridge penalty λ is selected from a fixed grid using the one-standard-error rule applied to the blocked-cross-validation root mean squared error (RMSE) curve. The same blocked-cross-validation principle is applied to the benchmark ridge specifications. The candidate grids and selected values are reported in the empirical design tables, rather than left implicit. This makes the model-selection design explicit and inspectable.

The distinction between w_t and ω_h is important. The observation weight w_t controls robustness across time episodes, whereas the horizon weight ω_h controls how the pathwise validation loss trades off short and long horizons. Table A2 also reports an ablation analysis evaluating whether the depth-weighted component, isotonic projection, horizon-weighted tuning, and horizon-specific penalty variants materially change performance.

3.5. Training-only regimes and robustness design

For regime robustness, we conduct a stratified evaluation across training-defined variance states. Variance regimes are defined using thresholds estimated from the training block only. The same frozen thresholds are then applied to calibration and test observations. This ensures that regime labels do not use future information. The empirical results report long-horizon RMSE separately for low-, normal-, and high-variance regimes.

The main empirical sample spans 2010–2025 and includes four liquid assets: GDX, GDXJ, XLE, and UUP. The long sample includes several volatility environments and both precious metal/equity linked and currency-related risk exposures. In addition to the main $H = 30$ design, Figure A3 reports horizon sensitivity for $H = 10, 20, 30$. Proxy robustness is assessed by repeating the FRV construction under alternative daily variance proxies, as summarized in Figure A1. These checks are treated as robustness and limitation analyses, not as evidence of broad superiority.

3.6. Benchmark specifications

The depth-weighted monotone functional forecast is compared with several benchmark classes. The first benchmark is a rolling historical FRV forecast, which uses the recent empirical average of past FRV paths and provides a nonparametric persistence-type baseline. The second benchmark is HAR ridge, a regularized multiresponse version of the heterogeneous autoregressive variance specification, motivated by the HAR-RV literature [4]. The third benchmark is FPCA-AR, which projects historical FRV paths onto leading functional principal components and forecasts the resulting

scores autoregressively. Figure A4 also reports GARCH(1,1)-based forecasts as a conventional volatility benchmark [29]. The main figures focus on the functional and HAR-type comparators, whereas the GARCH benchmark is shown in Figure A4.

The benchmark set is designed to separate three questions: whether the proposed method has lower loss than a simple rolling historical FRV baseline, whether it remains competitive with a HAR-type variance benchmark, and whether the functional representation adds information beyond a low-dimensional FPCA-AR forecast. The empirical interpretation is correspondingly cautious. The proposed model is evaluated as a leakage-disciplined, shape-aware functional enhancement within this benchmark set.

3.7. One-sided block-calibrated conformal envelopes

Uncertainty is summarized through upper one-sided block-calibrated conformal bands. Because financial time series are serially dependent, the conformal bands are not presented as exact finite-sample distribution-free intervals under arbitrary dependence. Exact split-conformal validity requires exchangeability assumptions that are not literally satisfied in this setting [14, 15, 30]. Accordingly, we interpret the bands as empirical block-calibrated risk envelopes and report coverage sensitivity across block lengths and guard factors. This wording avoids overstating the formal validity of conformal inference under temporal dependence.

Horizon-specific residual scales are estimated using training information only. Let

$$r_{t,h}^{\text{train}} = y_{t,h} - \widehat{y}_{t,h}^{\text{iso}}, \qquad t \in \mathcal{I}_{\text{train}}.$$

Define

$$\widehat{\sigma}_h = \text{MADN}\{r_{t,h}^{\text{train}} : t \in \mathcal{I}_{\text{train}}\},$$

with the same numerical fallback rule as above. These scales are frozen before calibration. Thus, the calibration block is reserved for estimating the conformal quantile, not for tuning, scaling, or selecting model components.

On the calibration set $\mathcal{I}_{\text{cal}}$, compute standardized one-sided residuals

$$e_{t,h} = \max\left\{\frac{y_{t,h} - \widehat{y}_{t,h}^{\text{iso}}}{\widehat{\sigma}_h}, 0\right\}.$$

Partition the calibration indices into consecutive nonoverlapping blocks $\{\mathcal{B}_b\}$ of length B, and compute the block score

$$m_b = \max_{t \in \mathcal{B}_b} \max_{1 \le h \le H} e_{t,h}. \tag{3.8}$$

Let m be the number of finite block scores. For nominal miscoverage level α, define

$$k = \lceil (m+1)(1-\alpha) \rceil,$$

clipped to $\{1, \ldots, m\}$, and let $\widehat{q}_{1-\alpha}$ be the kth order statistic of $\{m_b\}_{b=1}^{m}$.

The empirical threshold is defined by Eq (3.9):

$$q_\alpha = \gamma \widehat{q}_{1-\alpha}, \qquad \gamma \ge 1. \tag{3.9}$$

The factor γ is not used to claim a new conformal validity theorem; rather, it is treated as a guard factor sensitivity parameter for examining the trade-off between empirical coverage and band width under dependence. Figure A5 reports $\gamma \in \{1.00, 1.05, \ldots, 1.25\}$, and Table A4 reports the unguarded conformal threshold ($\gamma = 1$).

For a test forecast origin $t^\star$, the upper one-sided band is

$$U_{t^\star,h}(\alpha) = \widehat{y}^{\,\mathrm{iso}}_{t^\star,h} + q_\alpha\, \widehat{\sigma}_h, \qquad h = 1, \ldots, H. \tag{3.10}$$

The band is uniform over horizons because a single block-max threshold controls the whole FRV path. A test block is counted as covered if

$$\max_{t \in \mathcal{B}^{\mathrm{test}}} \max_{1 \le h \le H} \{y_{t,h} - U_{t,h}(\alpha)\} \le 0.$$

The default block length corresponds to a trading-week interpretation, and robustness checks vary the block length.

3.8. Equal predictive accuracy and MCS-style robustness screening

In addition to descriptive RMSE and percent-reduction-in-MSE (PRE) summaries, the empirical analysis reports evidence for equal predictive accuracy. For each model M, define a date-level long-horizon loss

$$\ell_t(M) = \frac{1}{|S|} \sum_{h \in S} \left(y_{t,h} - \widehat{y}^{\,\mathrm{iso}}_{t,h}(M)\right)^2, \qquad S = \{20, \ldots, 30\}. \tag{3.11}$$

For pairwise comparisons between the proposed model P and a benchmark M_0, the loss differential is

$$d_t = \ell_t(M_0) - \ell_t(P).$$

Positive average d_t favors the proposed model. Diebold–Mariano (DM) tests are computed on the date-level sequence $\{d_t\}$ using heteroskedasticity- and autocorrelation-consistent standard errors to account for serial dependence and overlapping forward targets [31, 32]. The reported tests are interpreted as complementary evidence rather than as the sole basis for model ranking.

We also report a model confidence set (MCS)-style robustness screen based on the same date-level loss matrix. The screen follows the model confidence set logic of asking whether a subset of models can be retained as statistically close when forecast losses are similar, while avoiding a forced single-winner interpretation [33]. This is particularly relevant here because the proposed model, HAR ridge, and FPCA-AR can be close in some assets and regimes. Accordingly, the empirical discussion distinguishes lower loss relative to rolling historical FRV from more asset-dependent comparisons against HAR-type benchmarks.

3.9. Uncertainty in efficiency gains via moving-block bootstrap

To quantify uncertainty in efficiency gains under serial dependence, we apply a moving-block bootstrap to paired test loss differences. The estimand is the mean improvement in squared error, primarily over the long-horizon set $S = \{20, \ldots, 30\}$, and secondarily over all horizons.

Let

$$d_{t,h} = \left(y_{t,h} - \widehat{y}^{\,\mathrm{iso}}_{t,h}(\text{benchmark})\right)^2 - \left(y_{t,h} - \widehat{y}^{\,\mathrm{iso}}_{t,h}(\text{proposed})\right)^2.$$

The long-horizon mean gain is

$$\bar{d}_S = \frac{1}{|\mathcal{I}_{\text{test}}||S|} \sum_{t \in \mathcal{I}_{\text{test}}} \sum_{h \in S} d_{t,h}. \tag{3.12}$$

Bootstrap resampling is applied to the paired loss-difference sequence over forecast origins. Contiguous blocks of length B_{mbb} are sampled with replacement, concatenated, and truncated to the test length. For each bootstrap replicate, $\bar{d}_S$ is recomputed, and percentile intervals are reported as a dependence-aware uncertainty summary [26, 27]. The bootstrap intervals are used to summarize sampling uncertainty; the DM tests and MCS-style screen provide complementary equal predictive accuracy evidence.

3.10. Forecasting algorithm

Algorithm 1 summarizes the full procedure. The algorithm emphasizes information timing. Model tuning, scaling, depth construction, and regime definition are restricted to the training block; calibration is used only for conformal quantile estimation; and the test block is used only for final evaluation.

Algorithm 1 Leakage-disciplined FRV forecasting and evaluation procedure

Require: Daily price information, variance proxy v_t, horizon H, chronological training/calibration/test split, grids for (a, β, η) and λ, conformal block length B, guard factor γ, and bootstrap block length B_{mbb}.
Ensure: Monotone FRV forecasts, benchmark comparisons, conformal bands, and uncertainty summaries.
Construct FRV targets $y_{t,h} = \sum_{j=1}^{h} v_{t+j}$ for $h = 1, \ldots, H$.
Split forecast origins chronologically into training, calibration, and test blocks.
Define variance regimes using training information only and freeze the resulting thresholds.
Select hyperparameters within the training block using blocked cross-validation.
for all candidate (a, β, η) **do**
    Compute depth weights on the relevant training observations.
    Fit candidate ridge models over the λ-grid with an unpenalized intercept.
    Apply isotonic projection to fitted FRV paths.
    Evaluate candidates using the horizon-weighted validation loss.
end for
Select $(a, \beta, \eta, \lambda)$ using the blocked-CV criterion and the one-standard-error rule for λ.
Refit the final proposed model on the full training block and obtain monotone forecasts.
Fit benchmark models under the same chronological information set.
Estimate horizon scales $\widehat{\sigma}_h$ from training residuals only.
On the calibration block, compute standardized one-sided residuals and block-max scores.
Estimate the conformal threshold $q_\alpha = \gamma \widehat{q}_{1-\alpha}$.
On the test block, compute point-forecast losses, PRE, regime-specific losses, and one-sided block coverage.
Conduct DM tests, MCS-style screening, and moving-block-bootstrap analyses using date-level paired loss sequences.

4. Empirical study

4.1. Setup

4.1.1. Assets, data source, and variance proxy construction

The empirical design uses a fixed four-asset ETF panel: GDX, GDXJ, XLE, and UUP. The assets are selected before model evaluation to represent metal-related equity risk, energy sector equity risk, and U.S. dollar exposure, rather than being selected according to realized forecasting performance. Daily open, high, low, close, and adjusted close information is obtained from Yahoo Finance over the 2010–2025 sample window [34]. The main specification uses close-to-close log returns computed from adjusted prices and defines the daily variance proxy as $v_t = r_t^2$. Because daily squared returns are noisy relative to high-frequency realized measures, the analysis is repeated using OHLC-based daily variance

proxies, including Parkinson, Garman–Klass, and Rogers–Satchell measures; the corresponding proxy robustness evidence is summarized in Figure A1.

For each forecast origin t, the forward realized variance (FRV) target is the cumulative future variance proxy path

$$y_{t,h} = \sum_{j=1}^{h} v_{t+j}, \qquad h = 1, \ldots, H.$$

The main horizon is $H = 30$, corresponding to an approximately monthly trading risk window. Robustness checks for $H = 10$ and $H = 20$ are reported in Figure A3. The target construction uses strictly future variance proxies, while the predictors use information available at the forecast origin. Thus, the design respects the information timing required for out-of-sample forecasting.

4.1.2. Chronological split and evaluation protocol

All experiments use a chronological training/calibration/test design. The training block is used for model fitting, scaling, depth construction, regime-threshold estimation, and hyperparameter selection. The calibration block is used only for conformal threshold estimation. The test block is held out for final forecast evaluation. Table 1 reports the fixed panel sample design. The same split structure is used across the four assets to ensure comparability of test period results.

Table 1. Chronological empirical design and fixed panel sample split.

Asset	**Period**	H	**Forecast origin counts**			
			n_{total}	n_{train}	n_{cal}	n_{test}
GDX	2010–2025	30	3971	2382	794	795
GDXJ	2010–2025	30	3971	2382	794	795
XLE	2010–2025	30	3971	2382	794	795
UUP	2010–2025	30	3971	2382	794	795

Note: Counts denote usable forecast origins after FRV target construction and complete-case filtering. The common test period is 16 September 2022–14 November 2025.

4.1.3. Models and evaluation metrics

The proposed model is compared with three main benchmark classes in the body of the paper: rolling historical FRV, HAR ridge, and FPCA-AR. A GARCH(1,1)-based benchmark is retained in Figure A4 because its horizon profiles can operate on a different scale and visually dominate the main panels. This benchmark structure separates a simple historical curve comparison from HAR-type and functional competitors.

Accuracy is evaluated using horizonwise RMSE, aggregate MSE, long-horizon MSE over $h = 20, \ldots, 30$, and PRE. For a benchmark model M_0 and the proposed model P, the long-horizon PRE is

$$\text{PRE}_{20:30} = 100\left(1 - \frac{\text{MSE}_{20:30}(P)}{\text{MSE}_{20:30}(M_0)}\right).$$

Positive PRE indicates a lower long-horizon MSE for the proposed model relative to the benchmark. Equal predictive accuracy evidence is reported using Diebold–Mariano tests with

autocorrelation-robust standard errors and a model confidence set (MCS)-style robustness screen. Figures A1–A6 report the GARCH inclusive horizon profiles, horizon length sensitivity, proxy robustness, HAR-relative cumulative gains, depth channel sensitivity, and guard factor sensitivity. Tables A1–A7 report the selected full model comparison, Diebold–Mariano tests, MCS-style retained sets, ablation results, selected tuning settings, depth channel sensitivity, and numerical conformal coverage values.

4.2. Results

4.2.1. Main horizonwise accuracy profiles

Figure 2 reports horizonwise RMSE profiles for the fixed ETF panel under the main $H = 30$ design. The figure compares the proposed depth-weighted model with HAR ridge, FPCA-AR, and rolling historical FRV. The profiles are consistent with the main long-horizon evidence: rolling historical FRV is generally less accurate in the relevant cumulative risk region, whereas the ranking against HAR ridge and FPCA-AR varies by asset. This pattern motivates the cautious interpretation used throughout the empirical discussion.

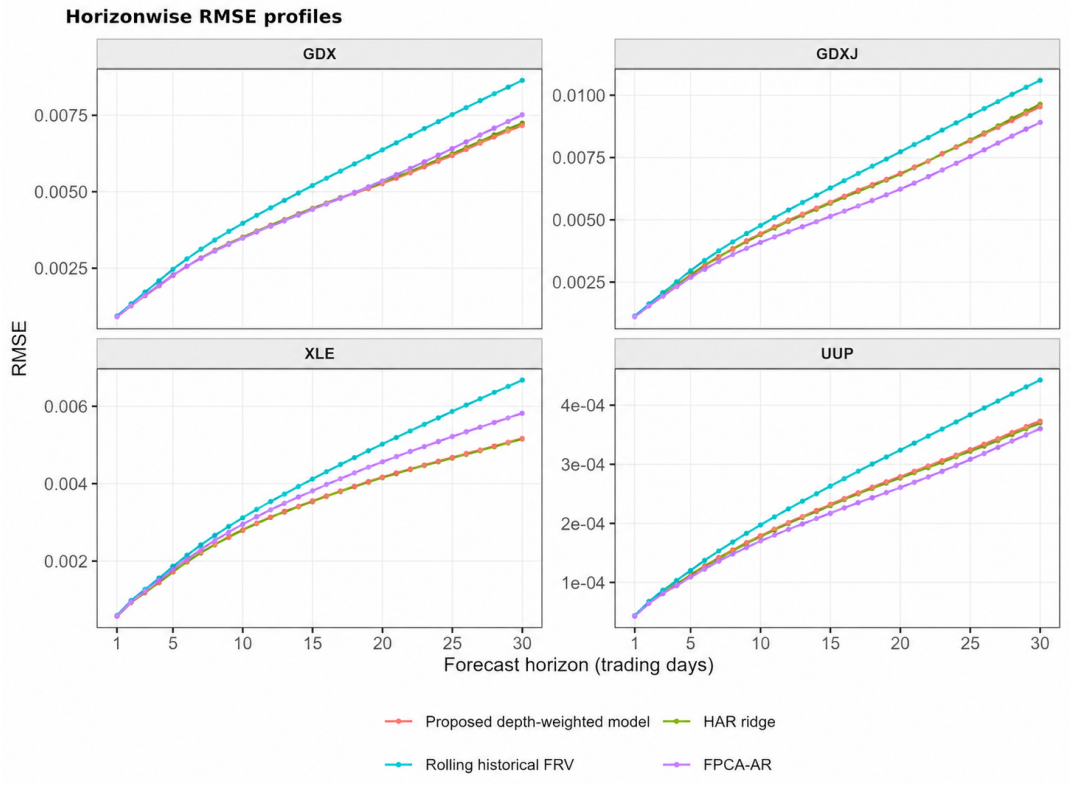

Figure 2. Horizonwise RMSE profiles for the fixed ETF panel. The figure compares the proposed depth-weighted model with HAR ridge, FPCA-AR, and rolling historical FRV under the main (H=30) design.

Table 2 summarizes the main long-horizon comparison. The proposed model has lower long-horizon MSE than rolling historical FRV for all four assets, with long-horizon PRE values ranging from approximately 20% to 37%. Relative to HAR ridge, the evidence is more modest and

asset-dependent: The proposed model is slightly favorable for GDX and GDXJ, whereas HAR ridge remains marginally stronger for XLE and UUP. This distinction keeps the interpretation from implying uniformly better performance across benchmark classes.

Table 2. Long-horizon accuracy gains of the proposed model over benchmark forecasts.

Asset	PRE relative to benchmark (%)		Main implication
	Rolling FRV	HAR ridge	
GDX	31.81	1.69	Improves on both benchmarks.
GDXJ	20.37	1.00	Improves on rolling FRV; close to HAR ridge.
XLE	36.54	−0.52	Large rolling-FRV gain; HAR ridge is marginally stronger.
UUP	28.03	−1.77	Improves on rolling FRV; HAR ridge remains stronger.

Note: PRE denotes percent reduction in MSE over $h = 20{:}30$. Positive values favor the proposed model relative to the stated benchmark. Full rankings and formal comparisons are reported in Tables A2 and A5–A7.

4.2.2. Cumulative long-horizon loss–gain relative to rolling historical FRV

Figure 3 plots the cumulative long-horizon loss–gain of the proposed model relative to rolling historical FRV. Positive values indicate lower cumulative loss for the proposed model. The figure provides a pathwise view of the same benchmark comparison summarized in Table 2: The proposed depth-weighted specification reduces cumulative long-horizon loss relative to rolling historical FRV across the fixed panel.

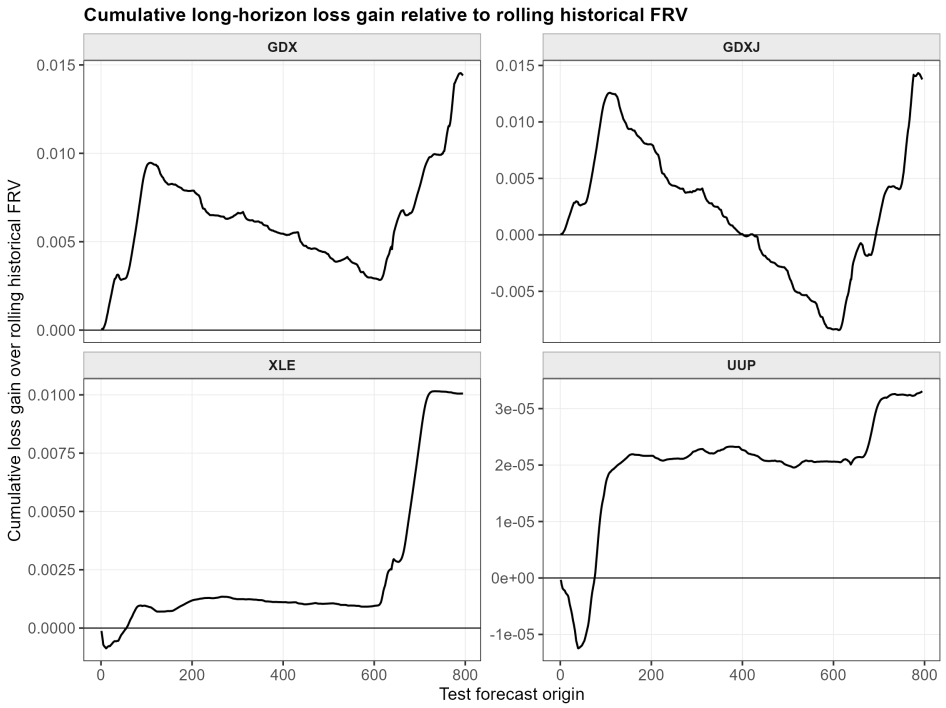

Figure 3. Cumulative long-horizon loss–gain relative to rolling historical FRV. Positive values favor the proposed depth-weighted model.

4.2.3. Regime-specific performance

Figure 4 reports regime-specific long-horizon RMSE using regimes defined from training information only. The purpose is to assess whether the forecasting evidence differs across variance states. The results show heterogeneous model rankings across regimes and assets. The proposed model is competitive in some high-variance regimes, particularly for the metal-related ETFs, whereas HAR ridge and FPCA-AR remain viable alternatives in other assets and states. Thus, the regime analysis supports a nuanced conclusion. Gains are present in several economically relevant states, but they are not uniform across all market environments.

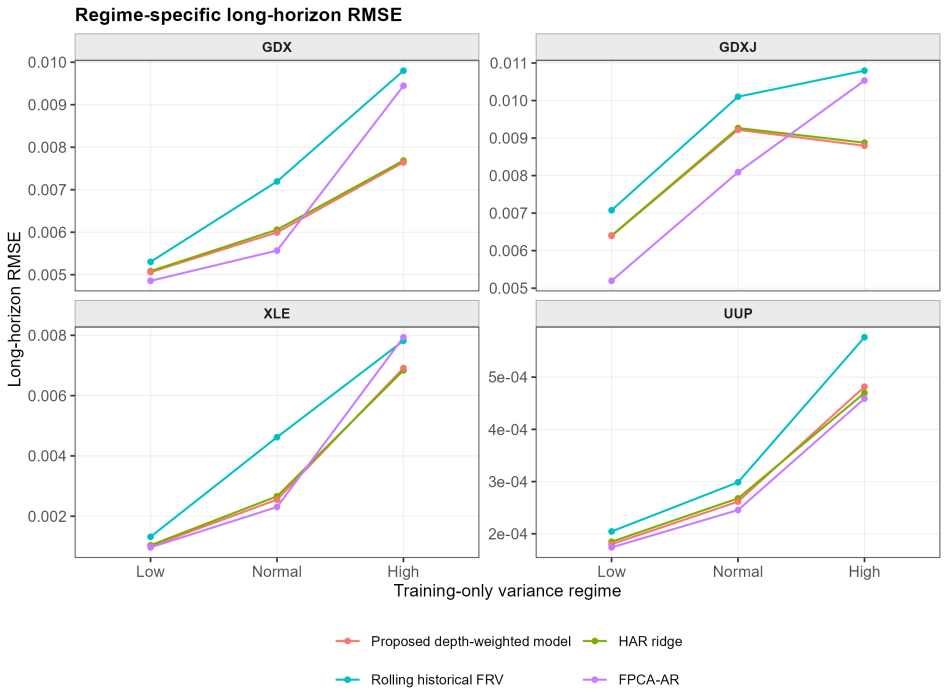

Figure 4. Regime-specific long-horizon RMSE. Regime thresholds are estimated using training information only and then applied to calibration and test observations.

4.2.4. Conformal coverage and uncertainty summaries

Table 3 summarizes the main conformal evidence for the proposed model under the unguarded threshold $\gamma = 1.00$. At $\alpha = 0.05$, empirical coverage is conservative or close to nominal for all four assets. At $\alpha = 0.10$, the evidence is more asset-dependent: GDX, GDXJ, and XLE remain above nominal coverage, whereas UUP falls below nominal coverage. The gap columns make this coverage behavior explicit. We therefore interpret the bands as empirical block-calibrated risk envelopes under temporal dependence rather than as exact distribution-free guarantees. Table A4 reports the corresponding thresholds and average widths, and Figure A5 reports guard factor sensitivity.

Overall, the empirical evidence supports the proposed method as a measured enhancement of daily FRV proxy forecasting. The most consistent finding is lower long-horizon loss relative to rolling historical FRV across the fixed panel. Comparisons with HAR ridge and FPCA-AR remain asset-dependent, as reflected in the main tables and the robustness evidence reported in Figures A2–A6 and

Tables A1–A7.

Table 3. Empirical block coverage of the one-sided conformal FRV bands.

Asset	$\alpha = 0.05$			$\alpha = 0.10$		
	Nominal	**Empirical**	**Gap**	**Nominal**	**Empirical**	**Gap**
GDX	0.950	1.000	+0.050	0.900	0.925	+0.025
GDXJ	0.950	1.000	+0.050	0.900	0.950	+0.050
XLE	0.950	1.000	+0.050	0.900	0.981	+0.081
UUP	0.950	0.956	+0.006	0.900	0.805	−0.095

Note: Coverage denotes empirical block coverage for the upper one-sided FRV band with block length $B = 5$ and unguarded threshold $\gamma = 1.00$. The gap is empirical coverage minus nominal coverage. Thresholds and average widths are reported in Table A4.

5. Conclusions and future directions

This paper develops a leakage-disciplined and shape-aware framework for forecasting forward realized variance (FRV) paths over a finite trading horizon. The proposed specification combines MRR, HDW based on shape and scale information, HW-BCV, IPP, and upper one-sided BCC bands. The design makes information timing explicit: model fitting, tuning, scaling, depth construction, and regime definition are restricted to the training block; calibration is used only for conformal threshold estimation; and the test block is reserved for final evaluation.

The empirical study uses a fixed four-asset ETF panel, GDX, GDXJ, XLE, and UUP, over the 2010–2025 period. The most consistent finding is that the proposed depth-weighted specification has lower long-horizon MSE than the rolling historical FRV benchmark across all four assets. This suggests that shape preservation, depth-based reweighting, and horizon-aware validation can be useful when the forecasting object is a cumulative variance path rather than a scalar one-step-ahead target.

At the same time, the evidence does not support a claim of uniformly better performance across all benchmarks. Comparisons with HAR ridge and FPCA-AR are asset-dependent. The proposed model is most favorable for GDX and remains close to HAR ridge for GDXJ, whereas HAR ridge and FPCA-AR remain viable alternatives for the remaining assets. The regime-specific results lead to a similar interpretation: the proposed method is useful in several high-variance and metal-related settings, but model rankings vary across assets and variance states.

The conformal component should also be interpreted cautiously. Because financial time series are serially dependent, the one-sided bands are presented as empirical block-calibrated risk envelopes rather than exact distribution-free guarantees. Coverage is conservative or close to nominal in several settings, especially at $\alpha = 0.05$, but the results are asset-dependent and UUP is more challenging at $\alpha = 0.10$. Figure A5 and Table A4 report guard factor and block size sensitivity, including the unguarded threshold $\gamma = 1.00$.

The study has several limitations. First, the FRV target is constructed from daily variance proxies, which are available and transparent but noisier than high-frequency realized measures. Second, the main horizon (H=30) is motivated by an approximately monthly risk window, although (H=10) and (H=20) checks are also reported. Third, the hybrid depth rule is deliberately simple and interpretable, and the depth channel sensitivity results show that shape-only, scale-only, and hybrid variants can behave differently by asset. Fourth, the evidence is based on a fixed ETF panel and should be extended

to other asset classes and market environments.

Future work may incorporate intraday realized measures or realized kernels, develop panel or multitask versions that borrow strength across related assets, compare richer functional-depth constructions, and allow horizon weights to adapt to market states or operational objectives. The conformal layer could also be refined through dependence-aware calibration, adaptive block selection, or covariate-shift adjustments.

Overall, the results support the proposed framework as a leakage-disciplined approach to daily FRV proxy forecasting. Its main empirical contribution is lower loss relative to rolling historical FRV for cumulative multiday variance paths, together with asset-dependent performance against HAR and FPCA-AR benchmarks. The method is therefore best viewed as a shape-aware forecasting layer for cumulative risk assessment.

Author contributions

Çağlar SÖZEN: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing–original draft, Writing–review & editing. **Onur ŞEYRANLIOĞLU**: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing–review & editing. **Arif ÇİLEK**: Methodology, Validation, Investigation, Data curation, Resources, Writing–review & editing. **Abdulmuttalip PİLATİN**: Conceptualization, Supervision, Validation, Project administration, Writing–review & editing. All authors read and approved the final version of the manuscript.

Use of Generative-AI tools declaration

The authors used ChatGPT (OpenAI) only for language polishing.

Data availability

The daily price data used in this study are publicly available from Yahoo Finance. Processed replication materials can be made available by the corresponding author upon reasonable request.

Funding

The authors declare that financial support was received for the research and/or publication of this article. This study was supported by the Recep Tayyip Erdoğan University Development Foundation under Grant No. 02026004016262.

Acknowledgments

The authors gratefully acknowledge the Recep Tayyip Erdoğan University Development Foundation for its support.

Conflict of interest

The authors declare no conflict of interests in this paper.

References

1. T. G. Andersen, T. Bollerslev, Answering the skeptics: Yes, standard volatility models do provide accurate forecasts, *Int. Econ. Rev.*, **39** (1998), 885–905. https://doi.org/10.2307/2527343
2. O. E. Barndorff-Nielsen, N. Shephard, Econometric analysis of realized volatility and its use in estimating stochastic volatility models, *J. R. Stat. Soc. B*, **64** (2002), 253–280. https://doi.org/10.1111/1467-9868.00336
3. T. G. Andersen, T. Bollerslev, F. X. Diebold, P. Labys, Modeling and forecasting realized volatility, *Econometrica*, **71** (2003), 579–625. https://doi.org/10.1111/1468-0262.00418
4. F. Corsi, A simple approximate long-memory model of realized volatility, *J. Financ. Economet.*, **7** (2009), 174–196. https://doi.org/10.1093/jjfinec/nbp001
5. A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12** (1970), 55–67. https://doi.org/10.1080/00401706.1970.10488634
6. R. Fraiman, G. Muniz, Trimmed means for functional data, *TEST,* **10** (2001), 419–440. https://doi.org/10.1007/BF02595706
7. S. López-Pintado, J. Romo, On the concept of depth for functional data, *J. Am. Stat. Assoc.*, **104** (2009), 718–734. https://doi.org/10.1198/jasa.2009.0108
8. S. Nagy, I. Gijbels, D. Hlubinka, Depth-based recognition of shape outlying functions, *J. Comput. Graph. Stat.*, **26** (2017), 883–893. https://doi.org/10.1080/10618600.2017.1336445
9. C. Bergmeir, J. M. Benítez, On the use of cross-validation for time series predictor evaluation, *Inform. Sci.*, **191** (2012), 192–213. https://doi.org/10.1016/j.ins.2011.12.028
10. C. Bergmeir, R. J. Hyndman, B. Koo, A note on the validity of cross-validation for evaluating autoregressive time series prediction, *Comput. Stat. Data Anal.*, **120** (2018), 70–83. https://doi.org/10.1016/j.csda.2017.11.003
11. M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, E. Silverman, An empirical distribution function for sampling with incomplete information, *Ann. Math. Statis.*, **26** (1955), 641–647. https://doi.org/10.1214/aoms/1177728423
12. R. E. Barlow, *Statistical inference under order restrictions: The theory and application of isotonic regression*, New York: John Wiley & Sons, 1972.
13. M. J. Best, N. Chakravarti, Active set algorithms for isotonic regression: A unifying framework, *Math. Program.*, **47** (1990), 425–439. https://doi.org/10.1007/BF01580873
14. V. Vovk, A. Gammerman, G. Shafer, *Algorithmic learning in a random world*, New York: Springer, 2005. https://doi.org/10.1007/b106715
15. J. Lei, L. Wasserman, Distribution-free prediction bands for nonparametric regression, *J. R. Stat. Soc. B*, **76** (2014), 71–96. https://doi.org/10.1111/rssb.12021
16. I. Gibbs, E. Candès, Adaptive conformal inference under distribution shift, *Adv. Neural Inform. Process. Syst.*, **34** (2021), 1660–1672.

17. M. Parkinson, The extreme value method for estimating the variance of the rate of return, *J. Bus.*, **53** (1980), 61–65. https://doi.org/10.1086/296071

18. M. B. Garman, M. J. Klass, On the estimation of security price volatilities from historical data, *J. Bus.*, **53** (1980), 67–78. https://doi.org/10.1086/296072

19. L. C. G. Rogers, S. E. Satchell, Estimating variance from high, low and closing prices, *Ann. Appl. Probab.*, **1** (1991), 504–512. https://doi.org/10.1214/aoap/1177005835

20. J. O. Ramsay, B. W. Silverman, *Functional data analysis*, 2 Eds., New York: Springer, 2005. https://doi.org/10.1007/b98888

21. L. Horváth, P. Kokoszka, *Inference for functional data with applications*, New York: Springer, 2012. https://doi.org/10.1007/978-1-4614-3655-3

22. J. S. Morris, Functional regression, *Annu. Rev. Stat. Appl.*, **2** (2015), 321–359. https://doi.org/10.1146/annurev-statistics-010814-020413

23. F. Ferraty, P. Vieu, *Nonparametric functional data analysis: Theory and practice*, New York: Springer, 2006. https://doi.org/10.1007/0-387-36620-2

24. A. Cuevas, A partial overview of the theory of statistics with functional data, *J. Stat. Plan. Infer.*, **147** (2014), 1–23. https://doi.org/10.1016/j.jspi.2013.04.002

25. Y. Romano, E. Patterson, E. Candès, Conformalized quantile regression, *Adv. Neural Inform. Proces. Syst.*, 2019.

26. H. R. Künsch, The jackknife and the bootstrap for general stationary observations, *Ann. Statist.*, **17** (1989), 1217–1241. https://doi.org/10.1214/aos/1176347265

27. S. N. Lahiri, *Resampling methods for dependent data*, New York: Springer, 2003. https://doi.org/10.1007/978-1-4757-3803-2

28. P. T. Reiss, R. T. Ogden, Functional principal component regression and functional partial least squares, *J. Am. Stat. Assoc.*, **102** (2007), 984–996. https://doi.org/10.1198/016214507000000527

29. T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, *J. Econometrics*, **31** (1986), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

30. J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, L. Wasserman, Distribution-free predictive inference for regression, *J. Am. Stat. Assoc.*, **113** (2018), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

31. F. X. Diebold, R. S. Mariano, Comparing predictive accuracy, *J. Bus. Econ. Stat.*, **13** (1995), 253–263. https://doi.org/10.1080/07350015.1995.10524599

32. W. K. Newey, K. D. West, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, **55** (1987), 703–708. https://doi.org/10.2307/1913610

33. P. R. Hansen, A. Lunde, J. M. Nason, The model confidence set, *Econometrica*, **79** (2011), 453–497. https://doi.org/10.3982/ECTA5771

34. *Historical price data (Adjusted Close and OHLC) for GDX, GDXJ, XLE, and UUP*, Yahoo Finance, 2025. Available from: `https://finance.yahoo.com/`.

Appendix

A. Additional robustness figures and tables

This appendix provides additional diagnostic and robustness evidence for the main empirical analysis. Figure A1 summarizes proxy robustness; Figure A2 examines depth channel sensitivity; Figure A3 reports horizon sensitivity; Figure A4 examines the full benchmark scope including GARCH(1,1); Figure A5 reports conformal guard factor sensitivity and Figure A6 reports HAR-relative cumulative gains.

Table A1 reports numerical depth channel sensitivity; Table A2 reports ablation results; Table A3 reports selected hyperparameters; Table A4 reports numerical conformal coverage values; Table A5 reports the full model comparison; Table A6 reports equal predictive accuracy evidence and Table A7 gives the MCS-style retained sets. These checks document robustness and diagnostic behavior rather than impose a single model ranking across all assets and settings.

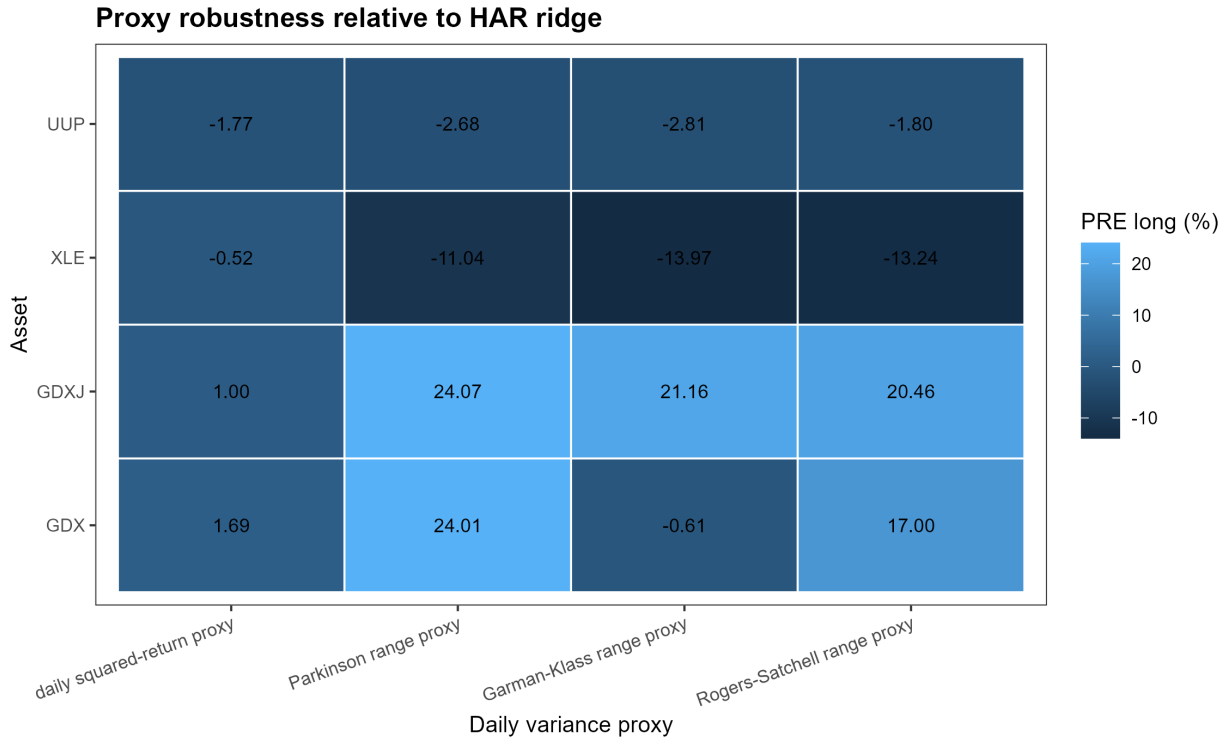

Figure A1. Proxy robustness heatmap. Entries summarize long-horizon performance relative to HAR ridge under alternative daily variance proxies. The evidence is interpreted as a robustness and limitation check rather than as a broad ranking result.

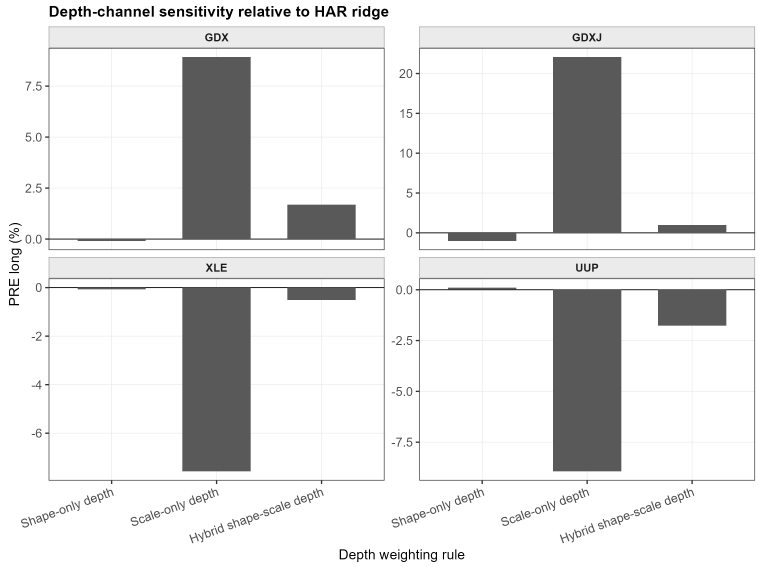

Figure A2. Depth channel sensitivity. The figure compares shape-only, scale-only, and hybrid depth specifications. The purpose is to evaluate sensitivity of the proposed weighting rule rather than to claim that the hybrid rule is uniformly optimal.

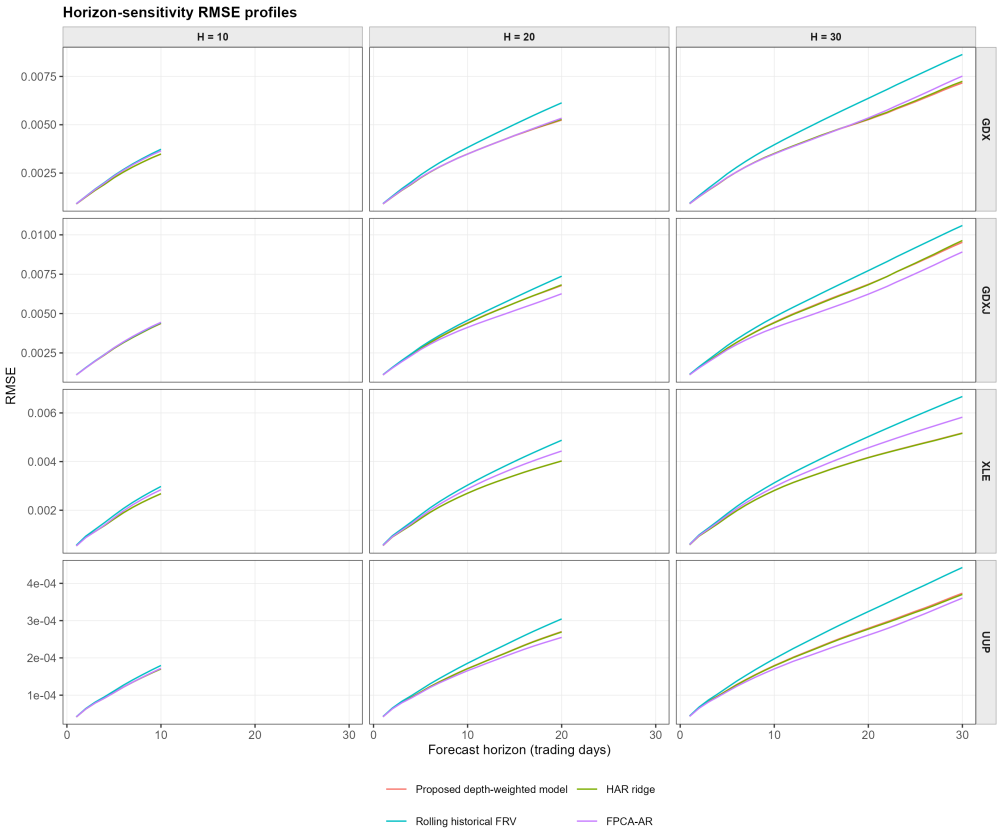

Figure A3. Horizon sensitivity of RMSE profiles for $H = 10, 20, 30$. The figure evaluates whether the main conclusions are specific to the monthly $H = 30$ horizon.

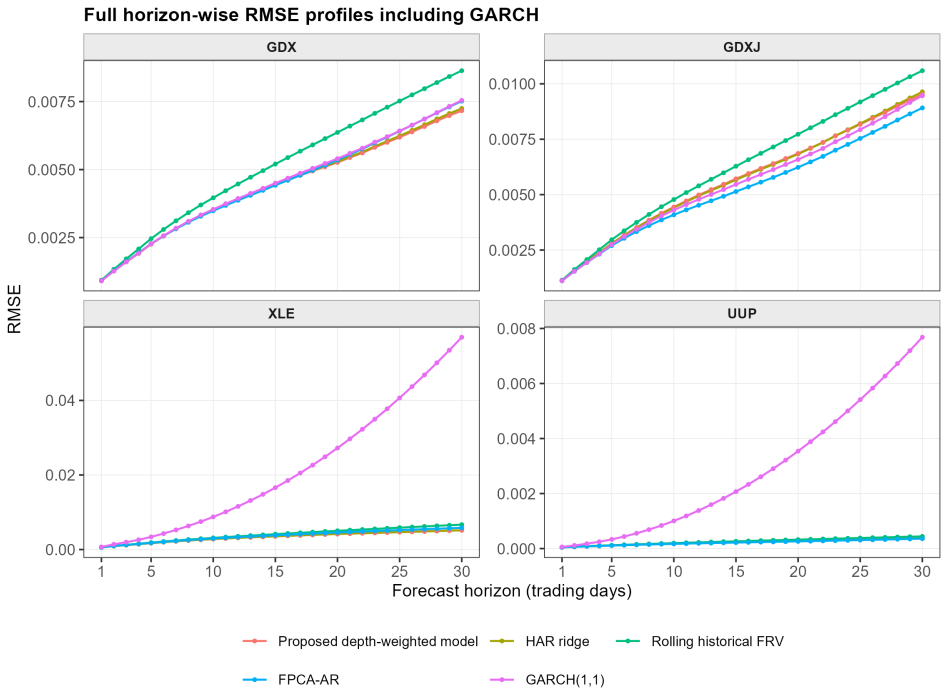

Figure A4. Full horizonwise RMSE profiles including GARCH(1,1). This figure reports the conventional GARCH benchmark alongside the main functional and HAR-type comparators.

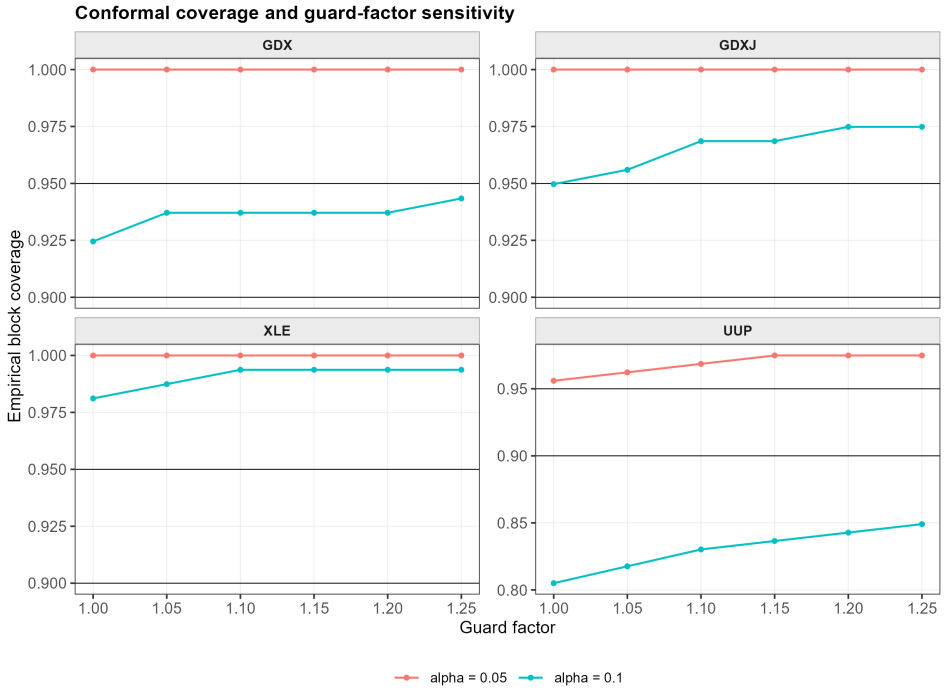

Figure A5. Conformal guard factor sensitivity. Coverage is reported as a function of the guard factor γ. The unguarded threshold $\gamma = 1.00$ is reported explicitly; larger values are interpreted as sensitivity checks under temporal dependence.

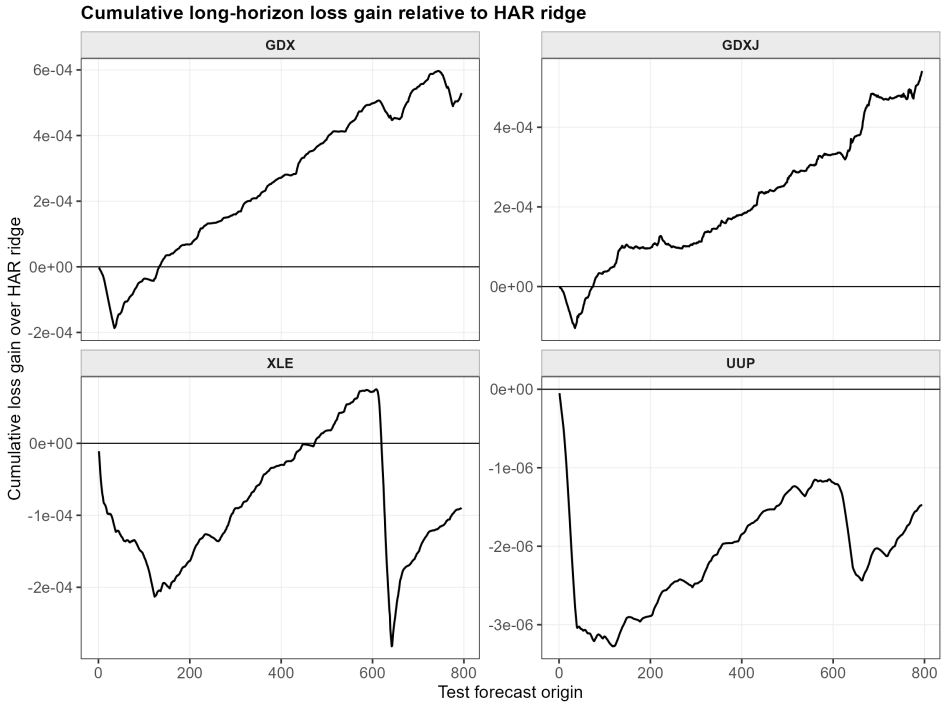

Figure A6. Cumulative long-horizon loss–gain relative to HAR ridge. Positive values favor the proposed model. The figure illustrates that gains relative to HAR ridge are asset-dependent.

Table A1. Depth channel sensitivity for long-horizon loss.

Asset	Scale-only depth		Shape-only depth		Hybrid shape–scale		Best channel
	$MSE_{20:30}$	**PRE**	$MSE_{20:30}$	**PRE**	$MSE_{20:30}$	**PRE**	
GDX	3.599×10^{-5}	8.92	3.955×10^{-5}	−0.09	3.885×10^{-5}	1.69	Scale-only
GDXJ	5.326×10^{-5}	22.06	6.903×10^{-5}	−1.01	6.766×10^{-5}	1.00	Scale-only
XLE	2.351×10^{-5}	−7.57	2.187×10^{-5}	−0.07	2.197×10^{-5}	−0.52	Shape-only
UUP	9.730×10^{-8}	7.29	1.053×10^{-7}	−0.29	1.068×10^{-7}	−1.77	Scale-only

Note: PRE denotes the percent reduction in long-horizon MSE relative to HAR ridge; positive values favor the listed depth specification. The best channel is determined by the lowest $MSE_{20:30}$ within each asset. The results show that the preferred depth channel is asset-dependent, so the hybrid rule is interpreted as a parsimonious baseline rather than a uniformly dominant specification.

Table A2. Ablation of shape enforcement, depth weighting, horizon-weighted tuning, and horizon-specific regularization.

Specification	Included components				Long-horizon performance		
	IPP	HDW	HW-BCV	H-λ	$\mathrm{MSE}_{20:30}$	PRE vs. HAR	Rank
Panel A: GDX							
Raw ridge	–	–	–	–	3.952×10^{-5}	0.00	4
HAR + IPP	✓	–	–	–	3.952×10^{-5}	0.00	4
HAR + HW-BCV	✓	–	✓	–	3.952×10^{-5}	0.00	4
Depth + IPP	✓	✓	–	–	3.885×10^{-5}	1.69	1
Proposed	✓	✓	✓	–	3.885×10^{-5}	1.69	1
Proposed + H-λ	✓	✓	✓	✓	3.885×10^{-5}	1.69	1
Panel B: GDXJ							
Raw ridge	–	–	–	–	6.834×10^{-5}	0.00	6
HAR + IPP	✓	–	–	–	6.834×10^{-5}	0.00	6
HAR + HW-BCV	✓	–	✓	–	6.834×10^{-5}	0.00	6
Depth + IPP	✓	✓	–	–	6.766×10^{-5}	1.00	3
Proposed	✓	✓	✓	–	6.766×10^{-5}	1.00	3
Proposed + H-λ	✓	✓	✓	✓	6.766×10^{-5}	1.00	3
Panel C: XLE							
Raw ridge	–	–	–	–	2.186×10^{-5}	0.00	1
HAR + IPP	✓	–	–	–	2.186×10^{-5}	0.00	1
HAR + HW-BCV	✓	–	✓	–	2.186×10^{-5}	0.00	1
Depth + IPP	✓	✓	–	–	2.197×10^{-5}	−0.52	5
Proposed	✓	✓	✓	–	2.197×10^{-5}	−0.52	5
Proposed + H-λ	✓	✓	✓	✓	2.197×10^{-5}	−0.52	5
Panel D: UUP							
Raw ridge	–	–	–	–	1.050×10^{-7}	0.00	2
HAR + IPP	✓	–	–	–	1.050×10^{-7}	0.00	2
HAR + HW-BCV	✓	–	✓	–	1.050×10^{-7}	0.00	2
Depth + IPP	✓	✓	–	–	1.068×10^{-7}	−1.77	6
Proposed	✓	✓	✓	–	1.068×10^{-7}	−1.77	6
Proposed + H-λ	✓	✓	✓	✓	1.068×10^{-7}	−1.77	6

Note: IPP denotes isotonic post-projection, HDW denotes hybrid depth weighting, HW-BCV denotes horizon-weighted blocked cross-validation, and H-λ denotes horizon-specific ridge regularization. PRE vs. HAR is the percent reduction in long-horizon MSE relative to HAR ridge; positive values favor the listed specification.

Table A3. Selected tuning parameters under the fixed chronological design.

Asset	**Selected validation parameters**			**Ridge penalty**
	a	β	η	λ_{ridge}
GDX	0.8	1.5	0.0	10^3
GDXJ	0.8	1.5	0.0	10^3
XLE	0.4	1.5	0.0	10^3
UUP	0.4	1.5	0.0	10^3

Note: All assets use the same chronological forecast origin split: $n_{\text{train}} = 2382$, $n_{\text{cal}} = 794$, and $n_{\text{test}} = 795$. Parameter a controls the shape–scale blend in HDW, while β and η define the horizon-weighted validation loss. λ_{ridge} is the selected ridge penalty.

Table A4. One-sided block-calibrated conformal coverage and band diagnostics.

Asset	$\alpha = 0.05$			$\alpha = 0.10$		
	Coverage	q	**Avg. width**	**Coverage**	q	**Avg. width**
GDX	1.000	32.019	9.300×10^{-2}	0.925	9.793	2.800×10^{-2}
GDXJ	1.000	40.552	1.580×10^{-1}	0.950	11.182	4.400×10^{-2}
XLE	1.000	35.332	1.060×10^{-1}	0.981	17.094	5.100×10^{-2}
UUP	0.956	21.355	2.000×10^{-3}	0.805	7.919	7.406×10^{-4}

Note: Results are for the proposed model under the unguarded one-sided conformal threshold $\gamma = 1.00$, using block length $B = 5$. Empirical coverage is computed over 159 test blocks for each asset. q denotes the empirical block-max calibration threshold, and Avg. width denotes the average upper-band addition on the FRV scale.

Table A5. Selected model comparison under the main $H = 30$ design.

Model	Loss		PRE (%)		Rank
	$MSE_{20:30}$	W-MSE	vs. rolling	vs. HAR	full set
Panel A: GDX					
Proposed full	3.885×10^{-5}	3.308×10^{-5}	31.81	1.69	1
HAR ridge + IPP	3.952×10^{-5}	3.360×10^{-5}	30.64	0.00	4
FPCA-AR	4.173×10^{-5}	3.519×10^{-5}	26.76	−5.58	8
GARCH(1,1)	4.199×10^{-5}	3.554×10^{-5}	26.30	−6.26	9
Rolling FRV	5.697×10^{-5}	4.794×10^{-5}	0.00	−44.17	10
Panel B: GDXJ					
FPCA-AR	5.774×10^{-5}	4.861×10^{-5}	32.04	15.51	1
GARCH(1,1)	6.438×10^{-5}	5.428×10^{-5}	24.23	5.79	2
Proposed full	6.766×10^{-5}	5.718×10^{-5}	20.37	1.00	3
HAR ridge + IPP	6.834×10^{-5}	5.761×10^{-5}	19.57	0.00	6
Rolling FRV	8.497×10^{-5}	7.132×10^{-5}	0.00	−24.33	11
Panel C: XLE					
HAR ridge + IPP	2.186×10^{-5}	1.887×10^{-5}	36.87	0.00	1
Proposed full	2.197×10^{-5}	1.896×10^{-5}	36.54	−0.52	5
FPCA-AR	2.729×10^{-5}	2.329×10^{-5}	21.16	−24.88	8
Rolling FRV	3.462×10^{-5}	2.921×10^{-5}	0.00	−58.40	9
GARCH(1,1)	2.000×10^{-3}	1.000×10^{-3}	−5068.35	−8086.65	11
Panel D: UUP					
FPCA-AR	9.666×10^{-8}	8.194×10^{-8}	34.88	7.93	1
HAR ridge + IPP	1.050×10^{-7}	8.919×10^{-8}	29.28	0.00	2
Proposed full	1.068×10^{-7}	9.071×10^{-8}	28.03	−1.77	6
Rolling FRV	1.484×10^{-7}	1.246×10^{-7}	0.00	−41.40	10
GARCH(1,1)	3.186×10^{-5}	2.480×10^{-5}	−21,364.70	−30,250.60	11

Note: $MSE_{20:30}$ is the average MSE over $h = 20{:}30$, and W-MSE is the horizon-weighted MSE used for model comparison. PRE denotes percent reduction in MSE relative to rolling historical FRV and HAR ridge; positive values favor the listed model. Ranks refer to the full candidate set, while the table reports selected representative models.

Table A6. Diebold–Mariano tests against the proposed model for long-horizon loss.

Benchmark	Mean diff.	DM	p-values		Favored model
			Raw	Holm	
Panel A: GDX					
Rolling FRV	1.812×10^{-5}	3.60	< 0.001	< 0.001	Proposed
HAR ridge + IPP	6.661×10^{-7}	3.86	< 0.001	< 0.001	Proposed
FPCA-AR	2.873×10^{-6}	0.95	0.170	0.511	Proposed
GARCH(1,1)	3.140×10^{-6}	1.17	0.122	0.487	Proposed
Panel B: GDXJ					
Rolling FRV	1.731×10^{-5}	1.77	0.039	0.231	Proposed
HAR ridge + IPP	6.807×10^{-7}	4.82	< 0.001	< 0.001	Proposed
FPCA-AR	-9.917×10^{-6}	−2.62	0.996	1.000	Benchmark
GARCH(1,1)	-3.275×10^{-6}	−0.57	0.716	1.000	Benchmark
Panel C: XLE					
Rolling FRV	1.265×10^{-5}	2.24	0.013	0.089	Proposed
HAR ridge + IPP	-1.149×10^{-7}	−2.03	0.979	1.000	Benchmark
FPCA-AR	5.315×10^{-6}	1.80	0.036	0.217	Proposed
GARCH(1,1)	1.978×10^{-3}	4.39	< 0.001	< 0.001	Proposed
Panel D: UUP					
Rolling FRV	4.159×10^{-8}	1.97	0.025	0.174	Proposed
HAR ridge + IPP	-1.861×10^{-9}	−1.62	0.947	1.000	Benchmark
FPCA-AR	-1.017×10^{-8}	−2.61	0.995	1.000	Benchmark
GARCH(1,1)	3.175×10^{-5}	3.21	0.001	0.009	Proposed

Note: Mean diff. denotes the average date-level long-horizon loss differential, defined as benchmark loss minus proposed model loss. Positive values favor the proposed model, whereas negative values favor the benchmark. DM denotes the Diebold–Mariano statistic computed with autocorrelation-robust standard errors. Holm values are adjusted within each asset-level comparison set.

Table A7. MCS-style retained model sets for long-horizon loss.

Asset	Best model	Best loss	Retained model set
GDX	Depth + IPP	3.885×10^{-5}	Depth + IPP; Proposed; Proposed + H-λ; FPCA-AR; GARCH(1,1).
GDXJ	FPCA-AR	5.774×10^{-5}	FPCA-AR.
XLE	Raw ridge	2.186×10^{-5}	Raw ridge; HAR + IPP; HAR + HW-BCV; Depth + IPP; Proposed; HAR H-λ; Proposed + H-λ; FPCA-AR.
UUP	FPCA-AR	9.666×10^{-8}	Raw ridge; HAR + IPP; HAR + HW-BCV; Depth + IPP; Proposed; HAR H-λ; Proposed + H-λ; Naive FRV; FPCA-AR.

Note: The table reports MCS-style retained sets based on the date-level long-horizon loss matrix. Retained models are not screened out as significantly worse within the robustness screen. The best model and retained set are computed over the full candidate set, whereas Table A5 reports selected representative models. The screen is used as robustness evidence and is not interpreted as a forced single-winner rule.