INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN ACCOUNTING, FINANCE AND MANAGEMENT SCIENCES

OPEN ACCESS JOURNAL

Vol 16, Issue 1, (2026) E-ISSN: 2225-8329

Volatility Shock Regimes and Risk-Aware Forecasting of Silver Futures Returns: A Rolling Out-of-Sample Evaluation with Conformal Intervals

Zhang Juan[1], Choo Wei Chong[2*], Yee Choy Leong[3], Lin Yihuan[4]

[1]Institute for Mathematical Research (INSPEM), Universiti Putra Malaysia, 43400, Serdang, Selangor Darul Ehsan, Malaysia, [2,3,4]School of Business and Economics, Universiti Putra Malaysia, 43400, Serdang, Selangor Darul Ehsan, Malaysia

*Corresponding Author Email: wcchoo@upm.edu.my

DOI Link: http://dx.doi.org/10.6007/IJARAFMS/v16-i1/27648

Published Online: 18 February 2026

Abstract

This study develops a leakage-free, regime-conditioned framework for forecasting silver futures returns and evaluating risk-aware performance under volatility shock regimes. Using daily SI=F data from January 2010 to January 2026, we construct log returns and a realized-volatility proxy $RV20$, and identify shock regimes via a rolling quantile threshold estimated strictly from past information, ensuring that regime classification remains executable out of sample. We benchmark a random-walk return forecast against a regularized linear ARX model and gradient-boosted trees and implement an ablation design to isolate the incremental contribution of monthly macro variables (oil, gold, and World Bank silver prices) that are aligned to daily observations using a one-month lag to prevent look-ahead bias. All models are evaluated through a rolling out-of-sample protocol with a frozen-hyperparameter strategy to preclude implicit test-time optimization. Results show that strong baselines remain difficult to outperform in RMSE, particularly during shock regimes, while directional accuracy exhibits horizon dependence, with linear dynamics more informative at short horizons and non-linear learners comparatively more stable at longer horizons. Predictive-accuracy tests indicate that macro augmentation does not deliver robust gains relative to strong benchmarks once information timing and estimation risk are controlled. To quantify uncertainty, we construct online rolling conformal prediction intervals and report regime-conditional calibration. Intervals widen materially during shocks, yet coverage deteriorates in stress states, consistent with distribution shift, implying that calibration behavior itself can serve as an operational trigger for hedging adjustment or capital preservation. Overall, the evidence emphasizes fair comparisons under identical information sets and highlights uncertainty quantification as a decision-critical output when return predictability is intrinsically limited.

Keywords: Silver Futures, Volatility Shock Regimes, Rolling Out-Of-Sample Evaluation, Xgboost, Conformal Prediction, Risk Management, Hedging

Introduction

Silver futures markets sit at the intersection of industrial demand, investment flows, and macro-financial uncertainty, making them central to corporate hedging, treasury risk control, and tactical allocation. Yet forecasting daily silver futures returns remains notoriously difficult. Returns exhibit time-varying volatility, heavy tails, and abrupt stress episodes in which the conditional distribution shifts rapidly, so average performance can obscure the regime-specific breakdowns that matter most for risk management. This motivates evaluation strategies that treat volatility as an organizing principle rather than a nuisance, consistent with foundational evidence on heteroskedasticity and volatility clustering (Engle, 1982; Andersen, Bollerslev, Diebold, & Labys, 2003).

The practical need is not simply to report marginal point-forecast gains, but to make forecasting usable as a monitoring input under strictly real-time information. For treasury and hedging decisions, the decision-relevant question is whether a forecasting system remains reliable when volatility shocks arrive, and whether uncertainty outputs can serve as objective triggers for escalation, such as tightening hedging rules, reducing exposure, or prioritizing capital preservation when distribution shift becomes pronounced. In daily horizons where predictability is often weak, operational value often shifts from better means to auditable error behavior and calibrated uncertainty, particularly during shock regimes.

Machine learning has renewed interest in nonlinear forecasting for commodity markets, with gradient-boosted trees widely adopted due to their flexibility and empirical performance in tabular settings (Chen & Guestrin, 2016). At the same time, the forecasting literature emphasizes that apparent gains often attenuate under strictly real-time protocols once information timing, repeated refitting, and strong baselines are fully respected (Giacomini & White, 2006). Two methodological concerns are especially salient for commodity return applications. First, mixed-frequency predictors are often incorporated in ways that risk implicit look-ahead, undermining fair comparisons. Second, uncertainty quantification is frequently treated as an auxiliary statistic, despite its central role in real-time risk control—particularly in precious metals, where macro conditions can be informative for volatility and risk but may translate only weakly into daily return predictability under conservative timing rules (Wang & Zhang, 2024; Ying & Luo, 2025; Wang et al., 2026).

Despite the growing use of machine learning in commodity forecasting, three gaps remain. First, mixed-frequency predictors are often used without fully enforcing real-time information timing, leaving leakage concerns unresolved. Second, stress-period performance is frequently averaged out, even though the most decision-relevant failures occur precisely during volatility shock regimes. Third, uncertainty quantification is typically reported as validation rather than operationalized as a deployable risk-governance layer, limiting practical value for treasury and hedging routines.

Accordingly, this study addresses three research questions. RQ1 asks whether gradient-boosted trees deliver robust rolling out-of-sample gains over strong benchmarks in forecasting silver futures returns, and whether any gains differ across shock versus non-shock

regimes. RQ2 examines whether monthly macro variables provide incremental predictive content once mixed-frequency alignment is strictly real-time and comparisons are conducted under matched information sets via an ablation design. RQ3 evaluates whether online conformal prediction intervals achieve reliable regime-conditional calibration, and whether interval behavior can be used as a trigger for hedging adjustment or capital preservation during identified shock regimes.

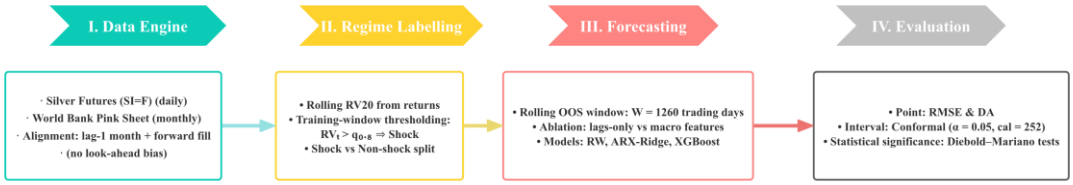

Figure 1 Framework for regime-conditioned forecasting and conformal risk calibration
Note: End-to-end pipeline linking leakage-free data alignment, RV20-based shock identification, rolling out-of-sample forecasting (W = 1260), and evaluation using RMSE/directional accuracy, Diebold–Mariano tests, and online conformal intervals (α = 0.05; cal = 252).

Figure 1 summarizes an operational pipeline that links leakage-free information timing to regime-aware evaluation in a single rolling out-of-sample design. Data Processing constructs daily log returns for silver futures and aligns monthly macro variables using a lag-one rule with forward filling to prevent look-ahead. Regime Identification labels volatility shocks via RV20 and a rolling training-window q0.8 threshold, ensuring the shock classifier remains executable out of sample. Model Training applies a W = 1260 rolling window and a matched-information ablation structure (lags-only versus lags-plus-macro) across RW, ARX–Ridge, and XGBoost. Evaluation reports RMSE and directional accuracy, Diebold–Mariano tests, and online rolling conformal intervals (α = 0.05; cal = 252), so interval behavior can be interpreted as a decision-critical trigger for hedging adjustment or capital preservation during shock regimes.

We address these questions with a regime-conditioned, leakage-free, and operationally executable framework. Monthly macro variables are aligned to daily observations using a one-month lag and forward filling to prevent look-ahead bias (Ghysels, Santa-Clara, & Valkanov, 2006). Volatility shock regimes are identified via a realized-volatility proxy and a rolling quantile threshold estimated strictly from past data, ensuring the regime classifier remains executable out of sample (Andersen et al., 2003; Hamilton & Susmel, 1994); Figure 2 provides a visual audit of this rule. We benchmark a random-walk return forecast against a regularized linear ARX model and gradient-boosted trees under matched information sets, evaluate multi-horizon forecasts (1-, 5-, and 20-day ahead), and adopt a frozen-hyperparameter strategy to preclude implicit test-time optimization (Chen & Guestrin, 2016). Statistical comparisons follow standard predictive-accuracy testing (Diebold & Mariano, 1995; Giacomini & White, 2006). To quantify uncertainty, we construct online rolling conformal prediction intervals and report regime-conditional coverage and width (Vovk, Gammerman, & Shafer, 2005; Angelopoulos & Bates, 2021; Gibbs & Candès, 2024); Figure 3 illustrates interval behavior in shock-dense windows.

The core message is deliberately conservative and therefore defensible. Under strict rolling evaluation, strong baselines remain difficult to outperform in RMSE, particularly during volatility shocks. Directional accuracy is horizon-dependent, suggesting that decision-relevant directional content can persist even when magnitude errors remain noise dominated. To mitigate metric cherry-picking, we report both squared- and absolute-loss criteria alongside directional accuracy, and we evaluate interval calibration via coverage and average width under each regime. Most importantly, conformal intervals widen during shocks but may still under-cover under stress, so calibration behavior itself becomes an objective, data-driven trigger for hedging adjustment or capital preservation when distribution shift is most pronounced (Gibbs & Candès, 2024).

The remainder of the paper proceeds as follows. The next section reviews work on volatility dynamics, mixed-frequency forecasting, and disciplined out-of-sample evaluation, and motivates uncertainty quantification for risk monitoring. The Method section details data construction, leakage-free alignment, regime identification, models, rolling evaluation, and conformal interval construction. The Results and Discussion section reports regime-conditioned point and interval performance and formal tests. The paper concludes with managerial implications and summarizes methodological contributions.

Literature Review

Volatility dynamics and the special case of silver

Financial returns exhibit volatility clustering and heavy tails, which motivates treating risk as time varying rather than constant (Engle, 1982; Bollerslev, 1986). Realized-volatility proxies operationalize this idea for monitoring and evaluation, and they are widely used to characterize time-varying risk in high-frequency settings (Andersen et al., 2003). Regime change is therefore central: volatility can shift discretely across states, as formalized in Markov-switching ARCH-type frameworks (Hamilton & Susmel, 1994), and regime-switching behavior has been documented for precious metals, including silver, under Markov-switching GARCH specifications (Naeem et al., 2019).

Silver is also a demanding forecasting target because it combines industrial-demand exposure with a precious-metal, stress-sensitive channel. This dual role helps explain why macro signals can appear economically relevant while remaining difficult to translate into stable daily return predictability under shifting volatility conditions, particularly when the main objective is risk management rather than alpha (Ying & Luo, 2025; Wang et al., 2026).

Mixed-frequency macro information, real-time alignment, and ragged-edge constraints

Mixed-frequency forecasting methods, including MIDAS-type designs, address the reality that predictors and targets arrive at different sampling rates (Ghysels et al., 2006). In real time, however, the core challenge is not simply “monthly versus daily” but ragged edges and data revisions. Macroeconomic and commodity indicators arrive with publication lags, are revised after first release, and update asynchronously across series, so naïve merges can embed information that was not available at the forecast origin (Foroni et al., 2015).

This implies that look-ahead bias is not only temporal leakage but also vintage leakage: backtests often use final revised data rather than the point-in-time values that decision-makers would have observed. The real-time macro literature shows that forecast conclusions

can change materially once vintage-consistent information sets are enforced (Croushore, 2002; Croushore, 2006). Against this backdrop, conservative alignment rules—such as lagging monthly series before mapping them to daily timestamps—are an audit-friendly way to minimize implicit timing assumptions, and they clarify why marginal macro gains in daily return forecasting can be fragile under strict protocols (Ghysels et al., 2006; Ying & Luo, 2025; Wang et al., 2026).

Rolling out-of-sample evidence, strong benchmarks, and fair comparisons

Return predictability is empirically weak and often episodic, which makes robust baselines difficult to beat on magnitude-based losses, consistent with weak-form efficiency arguments (Fama, 1970). This has elevated rolling out-of-sample designs and matched-information comparisons as credibility requirements, because they separate model capacity from incremental information content under repeated estimation (Giacomini & White, 2006). Inferential tools such as the Diebold–Mariano test provide disciplined evidence on loss differentials (Diebold & Mariano, 1995), but methodological discussions caution against interpreting significance tests as definitive rankings in unstable, dependent environments (Diebold, 2015). Recent commodity ML studies therefore emphasize leakage control, rolling evaluation, and ablation logic as prerequisites for credible claims rather than optional robustness checks (Wang & Zhang, 2024; Ye et al., 2025).

Machine learning, estimation risk, and the boundary between signal and model risk

Gradient-boosted decision trees are popular in commodity and precious-metals forecasting because they capture nonlinearities and interactions without imposing a parametric form (Chen & Guestrin, 2016). Precious-metals applications illustrate both promise and limits: XGBoost-based pipelines can be informative, especially when paired with interpretability tools, but performance is sensitive to state dependence and evaluation discipline (Jabeur et al., 2024; Wang & Zhang, 2024). When mixed-frequency macro covariates are added, estimation risk can dominate: higher-dimensional nonlinear learners may pay a variance cost under repeated refitting that offsets marginal information gains in noisy daily returns (Chen & Guestrin, 2016; Giacomini & White, 2006). This boundary is consistent with precious-metals evidence suggesting that macro information often expresses more robustly in volatility and uncertainty channels than in daily return levels (Ying & Luo, 2025; Wang et al., 2026).

Uncertainty quantification, conformal prediction, and regime-aware governance

When point-forecast gains are limited, decision value shifts toward calibrated uncertainty. Conformal prediction offers distribution-free predictive intervals that are less fragile than Gaussian bands under heavy tails, heteroskedasticity, and model misspecification (Vovk et al., 2005; Angelopoulos & Bates, 2021; Lei et al., 2018). The complication is that exchangeability is strained in time series, motivating rolling or adaptive conformal schemes under distribution shift (Gibbs & Candès, 2021) and sequential variants tailored to temporal dependence (Xu & Xie, 2021). In applied forecasting systems, this supports treating intervals as a governance layer: widening bands provide an interpretable warning signal, while systematic under-coverage during stress becomes a diagnostic that historical error behavior is no longer generalizing, prompting hedging adjustment or capital preservation responses (Angelopoulos & Bates, 2021; Gibbs & Candès, 2021; Luna et al., 2025).

Synthesis and Gap

The literature therefore points to a gap that is particularly relevant for silver futures. Volatility-regime research motivates stress conditioning, mixed-frequency work highlights ragged-edge and revision-driven leakage risks, forecast-evaluation studies stress rolling matched-information comparisons against strong baselines, and conformal inference provides a practical route to decision-critical uncertainty under misspecification and shift (Andersen et al., 2003; Foroni et al., 2015; Giacomini & White, 2006; Gibbs & Candès, 2021; Vovk et al., 2005). What remains underdeveloped is an integrated, leakage-disciplined framework that simultaneously (i) labels volatility shocks with an out-of-sample executable rule, (ii) isolates macro incremental value via ablation under fair information sets, and (iii) evaluates interval calibration conditional on regimes so uncertainty becomes an operational diagnostic rather than an auxiliary statistic.

Method

Research Design Overview

The study evaluates risk-aware return forecasting for silver futures under volatility shock regimes using a strict rolling out-of-sample OOS protocol. Two principles guide the design. First, predictors are aligned so that information used at each forecast origin is observable at that time, avoiding look-ahead bias. Second, models are compared under matched information sets, enabling an ablation interpretation that separates model capacity from incremental information content under repeated estimation (Giacomini & White, 2006). We consider horizons $\mathrm{h} \in 1,5,20$ and report performance for the full sample and separately for shock and non-shock regimes, consistent with recent commodity-forecasting evidence stressing rolling evaluation and leakage control in machine-learning applications (Wang & Zhang, 2024; Ye et al., 2025).

Data sources, returns, and mixed-frequency alignment

Daily silver futures prices are obtained from Yahoo Finance SI=F from 4 January 2010 to 29 January 2026. Let $\mathrm{P_t}$ denote the daily closing price and define continuously compounded percentage returns as

$$\mathrm{r_t} = 100(\log \mathrm{P_t} - \log \mathrm{P_{t-1}}).$$

Monthly macro-financial predictors are drawn from the World Bank silver spot price series. All monthly macro predictors enter the daily forecasting design only after a conservative lag-one alignment, so that information at each forecast origin is strictly available in real time. To prevent look-ahead bias when combining monthly and daily data, we apply a conservative lag-one alignment with forward filling: for any day t in month $\mathrm{m(t)}$, the daily macro predictor is

$$\mathrm{Z_t} = \mathrm{Z_{m(t)-1}}.$$

This alignment follows mixed-frequency forecasting discipline (Ghysels et al., 2006) and is consistent with recent precious-metals evidence that macro signals often manifest more robustly through volatility and uncertainty channels than through daily return levels under real-time constraints (Ying & Luo, 2025; Wang et al., 2026). Data cleaning and standardisation are implemented via a deterministic pipeline; implementation details are provided in the replication materials.

Targets, predictors, and information sets

For each horizon $h \in 1,5,20$, the target is $y_t^{(h)} = r_{t+h}$. Predictors are organised into two nested information sets. The endogenous block includes $r_{t-1}, \ldots, r_{t-10}$, a 20-day realised-volatility proxy,

$$RV20_t = \sqrt{\frac{1}{20}\sum_{i=0}^{19} r_{t-i}^2},$$

and 20-day rolling mean, minimum, and maximum of returns. The exogenous block contains the lag-aligned macro series oil, $gold$, and $silver_wb$. The first information set uses only endogenous predictors; the second augments them with macro predictors, enabling incremental-information comparisons under a fixed protocol (Andersen et al., 2003; Wang & Zhang, 2024).

Volatility shock regime identification

Shock regimes are defined using $RV20_t$ and a rolling threshold computed strictly from past information. With rolling window $W = 1260$ and quantile $q = 0.8$,

$$thr_t = Q_q(RV20_{t-W}, \ldots, RV20_{t-1}), \quad Shock_t = 1(RV20_t > thr_t).$$

Because thr_t uses only pre-t observations, regime labels are executable OOS and do not leak future volatility information. The approach aligns with regime-switching views of volatility dynamics (Hamilton & Susmel, 1994) and recent evidence on heterogeneous metal-market volatility motivating regime-aware evaluation (Bastianin et al., 2025). Figure 2 visualises the price, $RV20$, the rolling threshold, and the resulting shock intervals.

Forecasting models and rolling OOS protocol

We benchmark flexible learners against robust baselines under matched information sets. The random-walk-on-returns benchmark predicts zero. The linear benchmark is ARX–Ridge with fixed $\alpha = 1.0$ and within-window standardisation, adopting shrinkage to reduce estimation variance under repeated fitting (Hoerl & Kennard, 1970). Nonlinear models are gradient-boosted trees (XGBoost) estimated as XGB-lags and XGB-lags+macro (Chen & Guestrin, 2016). To strengthen academic defensibility, we implement a frozen-hyperparameter strategy during rolling OOS: hyperparameters are calibrated once on an initial training segment and held constant thereafter, precluding implicit test-time optimisation.

All models are evaluated in a fixed-length rolling scheme with training window $W = 1260$: at each origin t, models are fit on $[t - W, t - 1]$ and produce h-step-ahead forecasts. For XGBoost, early stopping uses an internal, time-ordered validation split (last 252 observations within the training window), preserving temporal ordering and avoiding post-origin information (Ye et al., 2025). The end-to-end pipeline—from data alignment to interval construction—is deterministic and computationally feasible for near-real-time monitoring in treasury and risk-management settings.

Evaluation metrics, DM tests, and conformal intervals

Point forecasts are evaluated using RMSE and MAE. Directional accuracy is computed as the fraction of non-negligible forecasts whose sign matches realised returns, excluding near-zero

forecasts from the denominator to avoid mechanical inflation in low-signal environments. Statistical comparisons use Diebold–Mariano tests under squared-error and absolute-error loss (Diebold & Mariano, 1995).

Uncertainty is quantified via online rolling conformal prediction. Let the nonconformity score be

$$s_t = |y_t - \hat{y}_t|.$$

With calibration window 252 and miscoverage $\alpha = 0.05$, the interval radius is the $(1-\alpha)$-quantile of recent scores, producing symmetric intervals $[\hat{y}_t - q_t, \hat{y}_t + q_t]$. This follows distribution-free conformal inference (Vovk et al., 2005; Angelopoulos & Bates, 2021), while regime-conditioned coverage serves as a diagnostic for distribution shift under time-series dependence (Gibbs & Candès, 2021; Xu & Xie, 2021). Recent applied work further motivates conformal intervals as operational risk bands in management and hedging contexts (Luna et al., 2025).

Results and Discussion

Table 1

Descriptive statistics of silver futures daily returns by volatility regime (full sample, shock, and non-shock)

	N	Mean	Std. Dev.	Skewness	Kurtosis	1%	5%	Median	95%	99%	Jarque–Bera
Whole	2762	0.069	1.919	-0.247	8.859	-5.039	-2.820	0.065	3.050	5.604	3961.1***
Shock	487	0.266	3.208	-0.368	5.381	-10.151	-4.609	0.258	5.627	7.434	122.4***
Non-shock	2275	0.026	1.505	-0.162	5.179	-4.195	-2.401	0.036	2.513	4.131	457.0***

Notes: Jarque–Bera is reported as JB statistic with significance stars (*** p<0.01, ** p<0.05, * p<0.10).

Table 1 reports pronounced regime dependence under the rolling $\mathrm{RV20}$-based indicator. Dispersion and downside tail risk intensify in shocks: the return standard deviation increases from 1.505 (non-shock) to 3.208 (shock), and the 1% quantile shifts from −4.195 to −10.151. Jarque–Bera statistics reject Gaussianity across partitions, consistent with conditional heteroskedasticity and heavy-tailed innovations in financial returns (Engle, 1982; Bollerslev, 1986). Figure 2 provides a visual audit: $RV20$ clusters and the rolling quantile threshold adapts over time, generating contiguous shock episodes rather than isolated spikes. This regime proxy is consistent with regime-switching views of volatility dynamics (Hamilton & Susmel, 1994) and documented volatility-state shifts in silver returns under Markov-switching GARCH specifications (Naeem et al., 2019), while remaining operationally executable out of sample and compatible with the leakage-free rolling design.

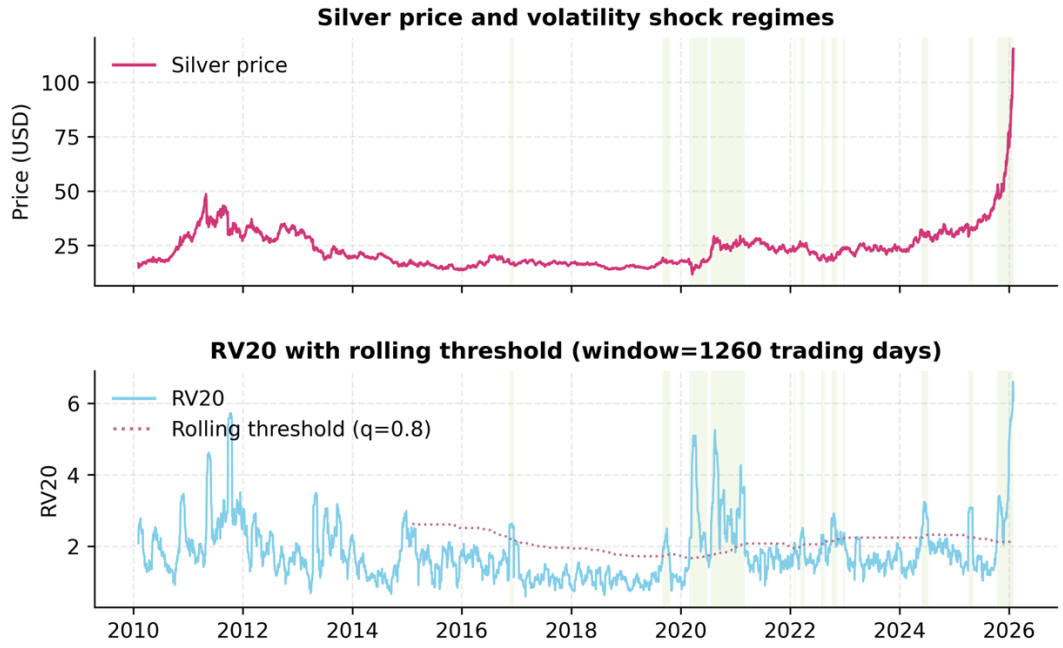

Figure 2 Conformal prediction intervals under volatility shock regimes
Notes: Realized returns, point forecasts, and the 95% online rolling conformal interval (cal = 252; α = 0.05) over a shock-dense window.

Regime-conditioned point-forecast performance: accuracy and directional content across horizons

Table 2 reports rolling OOS RMSE and directional accuracy for $(\mathrm{h} \in 1,5,20)$ by regime under matched information sets. Two patterns dominate. First, shocks are systematically harder: RMSE rises sharply for all models in the shock regime (e.g., at $h = 1$, RW RMSE increases from 1.595 to 3.007; XGB-lags from 1.614 to 3.049). Second, RMSE separation is small and unstable, with the random-walk benchmark remaining highly competitive. Under strict timing discipline, this is best read as a design-validity signal rather than a modeling failure, consistent with weak daily return predictability and elevated overfitting risk under repeated re-estimation (Fama, 1970; Giacomini & White, 2006).

Table 2
Rolling out-of-sample point-forecast performance across horizons and regimes: RMSE and directional accuracy (RW, ARX–Ridge, and XGBoost)

Panel A: h = 1 (N: Whole=2761, Shock=486, non-shock=2275)

Model	RMSE			DA		
	Whole	Shock	Non-shock	Whole	Shock	Non-shock
RW	1.921	3.007	1.595	-	-	-
ARX-Ridge	1.941	3.015	1.622	0.507	0.537	0.501
XGB-lags	1.945	3.049	1.614	0.493	0.516	0.487
XGB-lags+macro	1.948	3.046	1.620	0.485	0.523	0.477

Panel B: h = 5 (N: Whole=2757, Shock=482, non-shock=2275)

Model	RMSE			DA		

	Whole	Shock	Non-shock	Whole	Shock	Non-shock
RW	1.921	2.966	1.614	-	-	-
ARX-Ridge	1.938	3.002	1.625	0.495	0.500	0.494
XGB-lags	1.935	2.995	1.624	0.500	0.515	0.497
XGB-lags+macro	1.965	3.006	1.664	0.493	0.504	0.491
Panel C: h = 20 (N: Whole=2742, Shock=467, non-shock=2275)						
Model	**RMSE**			**DA**		
	Whole	**Shock**	**Non-shock**	**Whole**	**Shock**	**Non-shock**
RW	1.921	2.619	1.744	-	-	-
ARX-Ridge	1.932	2.636	1.753	0.509	0.488	0.513
XGB-lags	1.931	2.637	1.751	0.515	0.520	0.514
XGB-lags+macro	1.931	2.636	1.752	0.503	0.484	0.506

Notes: ARX-Ridge is a linear ARX model estimated with ridge regularisation using the same macro information set as XGB+macro. DA is directional accuracy; RW does not provide a directional forecast in returns and is shown as '-'. All RMSE and DA values are reported to three decimals. MAE is reported in the appendix.

Directional accuracy provides an economically interpretable complement. Table 2 shows clear horizon dependence: ARX–Ridge is relatively strongest at $h = 1$(0.507 full sample; 0.537 shock), whereas tree models become more competitive at longer horizons; at $h = 20$, XGB-lags exceeds ARX–Ridge in the shock regime (0.520 vs. 0.488). A plausible mechanism is that short-horizon direction is governed by simpler linear lag structure captured with low variance under shrinkage, while longer horizons allow nonlinear interactions and threshold effects that boosted trees can exploit more robustly (Chen & Guestrin, 2016; Giacomini & White, 2006). Hence, directional content can remain economically relevant even when magnitude accuracy is noise dominated, implying a more stable medium-horizon directional signal for tactical allocation when nonlinear state dependence becomes salient.

Robustness to loss functions: MAE results

Table 3

Rolling out-of-sample MAE by horizon and regime

Panel A: h = 1 (N: Whole=2761, Shock=486, non-shock=2275)			
Model	**MAE**		
	Whole	**Shock**	**Non-shock**
RW	1.304	2.084	1.137
ARX-Ridge	1.328	2.096	1.164
XGB-lags	1.321	2.109	1.153
XGB-lags+macro	1.327	2.108	1.161
Panel B: h = 5 (N: Whole=2757, Shock=482, Non-shock=2275)			
Model	**MAE**		
	Whole	**Shock**	**Non-shock**
RW	1.303	2.059	1.143
ARX-Ridge	1.324	2.118	1.156
XGB-lags	1.317	2.083	1.154
XGB-lags+macro	1.345	2.098	1.185

Panel C: h = 20 (N: Whole=2742, Shock=467, Non-shock=2275)			
Model	MAE		
	Whole	Shock	Non-shock
RW	1.304	1.784	1.206
ARX-Ridge	1.319	1.825	1.216
XGB-lags	1.314	1.818	1.211
XGB-lags+macro	1.319	1.825	1.215

Notes: Table reports mean absolute error (MAE) for each horizon and subsample. N is reported in the panel headers. All MAE values are reported to three decimals.

Table 3 reports MAE and corroborates the RMSE narrative: absolute errors rise materially in shocks for all models (e.g., RW MAE increases from 1.137 to 2.084 at $h = 1$) and stable ranking differences remain compressed. Thus, conclusions do not hinge on squared-loss sensitivity to extremes.

Loss-differential inference and information-set ablation: DM evidence on macro augmentation

Table 4
Diebold–Mariano evidence on relative predictive accuracy in rolling out-of-sample forecasts (h = 1, 5, 20)

Panel A: h = 1 (baseline: XGB+macro)

Compare	Loss	DM_stat	p_value	ΔLoss (A−B)	N
vs ARX-Ridge	SE	0.480	0.631	0.0265	2761
vs ARX-Ridge	AE	-0.067	0.946	-0.0005	2761
vs RW	SE	3.538***	<0.001	0.1065	2761
vs RW	AE	4.655***	<0.001	0.0233	2761

Panel B: h = 5 (baseline: XGB+macro)

Compare	Loss	DM_stat	p_value	ΔLoss (A−B)	N
vs ARX-Ridge	SE	2.421**	0.016	0.1073	2757
vs ARX-Ridge	AE	2.360**	0.018	0.0207	2757
vs RW	SE	4.632***	<0.001	0.1746	2757
vs RW	AE	5.210***	<0.001	0.0416	2757

Panel C: h = 20 (baseline: XGB+macro)

Compare	Loss	DM_stat	p_value	ΔLoss (A−B)	N
vs ARX-Ridge	SE	-0.184	0.854	-0.0044	2742
vs ARX-Ridge	AE	-0.102	0.919	-0.0006	2742
vs RW	SE	1.827*	0.068	0.0374	2742
vs RW	AE	2.883***	0.004	0.0145	2742

Notes: DM uses loss differentials d_t = L_A,t − L_B,t with A = XGB+macro and B = comparator. ΔLoss(A−B) is mean(d_t); ΔLoss>0 implies the comparator has lower average loss. Stars on DM_stat: *** p<0.01, ** p<0.05, * p<0.10.

Table 4 reports Diebold–Mariano tests under squared-error and absolute-error loss (Diebold & Mariano, 1995), comparing the macro-augmented boosted model against RW and ARX–

Ridge. The pattern is consistent: macro augmentation does not deliver robust incremental gains under leakage-free mixed-frequency alignment and in several cases significantly worsens performance, particularly at short and medium horizons (p-values below 0.001 against RW at $h = 1$ and $h = 5$). Relative to ARX–Ridge, the clearest deterioration appears at $h = 5$ (p-values around 0.02), while differences at $h = 1$ and $h = 20$ are not statistically significant. This supports an ablation interpretation: once timing discipline is enforced, monthly macro covariates do not provide a stable marginal signal for daily silver returns and can raise estimation variance in flexible learners, consistent with a bias–variance tradeoff under repeated rolling re-fitting (Chen & Guestrin, 2016; Giacomini & White, 2006). This should be interpreted as a boundary condition rather than a rejection of macro relevance: mixed-frequency forecasting emphasizes that low-frequency predictors contribute depending on aggregation and weighting (Ghysels et al., 2006), and recent precious-metals evidence suggests macro uncertainty often maps more naturally into volatility and uncertainty channels than daily return levels under conservative real-time alignment (Ying & Luo, 2025; Wang et al., 2026). Forecast-comparison tests are best treated as disciplined evidence rather than definitive "model truth," especially in rolling environments (Diebold, 2015).

Given the limited and unstable gains in point forecasts, we next shift attention from first-order accuracy to second-order risk characterization—whether uncertainty can be quantified in a regime-sensitive and operationally useful way via conformal prediction.

Regime-sensitive uncertainty quantification: conformal intervals as risk envelopes and shift diagnostics

Table 5

Regime-conditioned calibration of 95% online rolling conformal intervals: empirical coverage and average width (h = 1, 5, 20)

Panel A: h = 1 (N: Whole=2711, Shock=486, Non-shock=2225)

Model	Coverage			AvgWidth		
	Whole	Shock	Non-shock	Whole	Shock	Non-shock
XGB+macro	0.934	0.860	0.950	7.408	9.229	7.011
ARX-Ridge	0.937	0.870	0.952	7.414	9.131	7.039

Panel B: h = 5 (N: Whole=2707, Shock=482, Non-shock=2225)

Model	Coverage			AvgWidth		
	Whole	Shock	Non-shock	Whole	Shock	Non-shock
XGB+macro	0.937	0.882	0.949	7.490	9.442	7.067
ARX-Ridge	0.937	0.880	0.949	7.451	9.362	7.037

Panel C: h = 20 (N: Whole=2692, Shock=467, Non-shock=2225)

Model	Coverage			AvgWidth		
	Whole	Shock	Non-shock	Whole	Shock	Non-shock
XGB+macro	0.936	0.919	0.940	7.361	9.483	6.916
ARX-Ridge	0.938	0.912	0.943	7.337	9.398	6.904

Notes: Coverage is the empirical proportion of realised returns falling within the conformal interval; AvgWidth is the mean interval width (upper−lower) in return percentage points. Intervals are constructed using rolling calibration on past absolute errors.

Table 5 reports empirical coverage and average width for 95% online rolling conformal intervals, and Figure 3 provides a diagnostic zoom. Two results are central. First, intervals widen substantially in shocks, producing an interpretable risk envelope even when point forecasts show limited gains; for XGB+macro, average width increases from 7.011 (non-shock) to 9.229 (shock) at $h = 1$ with similar widening at $h = 5$ and $h = 20$. This aligns with applied use where conformal bands serve as operational uncertainty ranges (Luna et al., 2025) and with conformal inference as a distribution-free uncertainty layer (Angelopoulos & Bates, 2021; Vovk et al., 2005). Second, coverage deteriorates materially in shocks despite widening—coverage drops from 0.950 to 0.860 at $h = 1$ and from 0.949 to 0.882 at $h = 5$, remaining below nominal at $h = 20$(0.919)—indicating distribution shift and tail stress that outpace finite-window calibration. This pattern is consistent with conformal inference under shifting environments and time-series dependence, where exchangeability violations can induce under-coverage unless calibration becomes more adaptive or regime-aware (Gibbs & Candès, 2021; Angelopoulos & Bates, 2021; Xu & Xie, 2021). Consequently, under-coverage in shocks serves as a fail-safe diagnostic that identifies temporal boundaries where historical error patterns cease to generalize, providing a quantitative trigger for more conservative safety margins in risk governance.

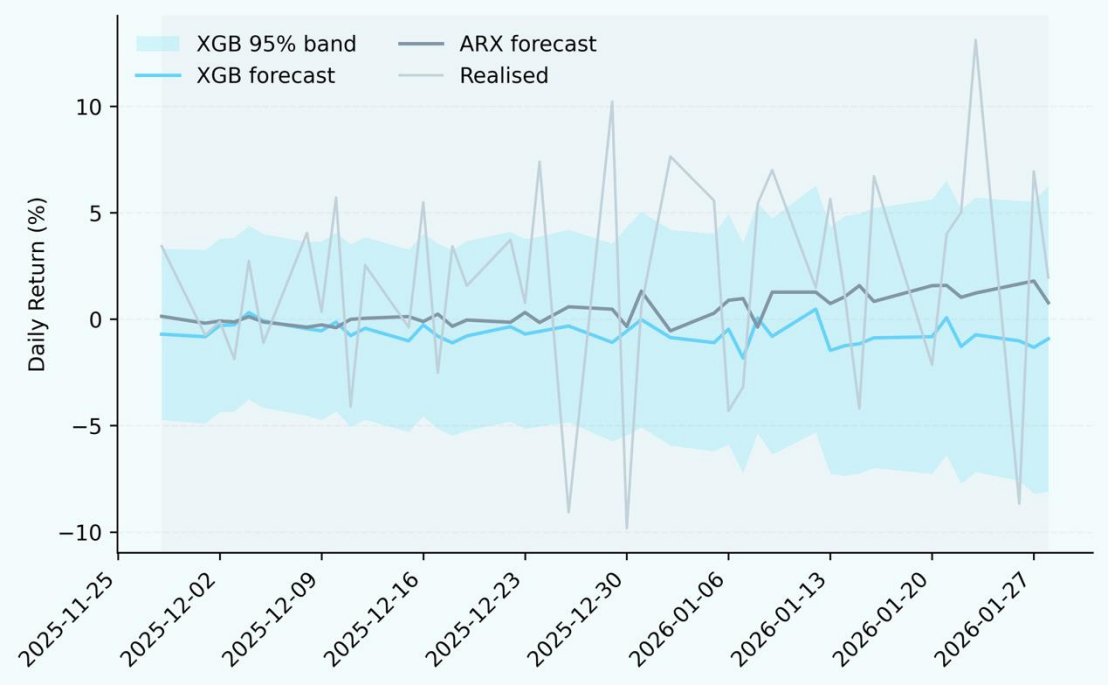

Figure 3 Rolling conformal prediction intervals in a shock-dense window

Note. The figure plots realised daily silver futures returns against rolling out-of-sample point forecasts from ARX–Ridge and XGBoost. The shaded band denotes the 95% online rolling conformal prediction interval around the XGBoost forecast, constructed with a calibration window of 252 trading days and nominal miscoverage level (\alpha = 0.05).

Conclusion

We conclude by summarizing what the rolling out-of-sample evidence implies for forecasting silver futures returns under volatility shock regimes. The benchmark comparisons align with a central lesson in financial forecasting: once strict information timing, repeated re-estimation, and leakage-free mixed-frequency alignment are enforced, incremental gains in return point forecasts are typically small and unstable, especially in turbulent episodes (Tang et al., 2022; Giacomini & White, 2006). Shock-conditioned diagnostics show that forecast errors rise sharply across models during volatility shocks, consistent with stress regimes compressing exploitable structure while amplifying model risk and parameter uncertainty (Hamilton & Susmel, 1994; Andersen et al., 2003). This is not a modelling "failure" but a credibility check: the absence of systematic dominance over robust baselines under a disciplined rolling protocol is consistent with episodic predictability and time-varying efficiency (Bock & Geissel, 2024).

Directional content can remain economically informative even when magnitude accuracy is noise dominated. Directional accuracy is horizon-dependent: ARX–Ridge is comparatively more reliable at short horizons, while the tree-based learner is relatively stronger at longer horizons, consistent with linear short-horizon dynamics and more pronounced nonlinear state dependence over longer windows (Chen & Guestrin, 2016; Wang & Zhang, 2024). Practically, linear signals may support immediate positioning, whereas nonlinear interactions may yield a more stable medium-horizon directional signal for tactical allocation, even if RMSE improvements remain modest (Tang et al., 2022; Wang & Zhang, 2024).

Macro augmentation is informative precisely because it is mixed. Although mixed-frequency macro information is meaningful for volatility and uncertainty in precious metals, translating low-frequency signals into stable gains for high-frequency return prediction is fragile under conservative alignment (Ying & Luo, 2025; Wang et al., 2026). Occasional degradation of the macro-augmented boosted model relative to parsimonious benchmarks is consistent with estimation risk and the bias–variance tradeoff: marginal information gains from monthly covariates can be outweighed by increased estimation variance in repeatedly refit nonlinear architectures (Hastie et al., 2009; Tang et al., 2022). This delimits where macro signals are most usable—often through risk and uncertainty channels rather than daily return levels (Ying & Luo, 2025; Wang et al., 2026).

Because point forecasts struggle to deliver large, robust gains, the framework's main operational value is uncertainty quantification. Conformal prediction intervals widen during shock regimes, yielding an interpretable risk envelope for monitoring and escalation (Vovk et al., 2005; Angelopoulos & Bates, 2021). Yet coverage deteriorates in shocks despite wider bands, consistent with distribution shift outpacing finite-window calibration (Gibbs & Candès, 2024; Gibbs & Candès, 2021). For corporate treasurers and risk managers, this implies a strategy shift during identified shocks: pivot from seeking alpha via aggressive directional positioning to beta-focused risk containment, using band widening as an objective trigger to reduce leverage, tighten stop-loss limits, or temporarily revert to benchmark-driven hedging rules until coverage stabilizes (Tang et al., 2022; Luna et al., 2026). The framework is therefore governance-oriented: it operationalizes stress recognition and uncertainty escalation under a deterministic rolling pipeline even when return-level predictability is intrinsically limited.

Contributions

This study offers a defensible, risk-aware forecasting pipeline for silver futures returns designed to withstand common reviewer concerns about leakage, evaluation optimism, and opportunistic tuning. It integrates leakage-free mixed-frequency alignment for monthly macro signals, a volatility shock proxy defined by rolling thresholds computed strictly from past information, and online rolling conformal intervals for sequential uncertainty quantification (Ghysels et al., 2006; Andersen et al., 2003; Vovk et al., 2005). Each component is executable at the forecast origin without hindsight, making the system replicable and audit ready.

A second contribution is a disciplined comparison design that separates information-set effects from model-class effects. The study evaluates a robust baseline, a regularized linear ARX benchmark, and gradient-boosted trees under the same rolling protocol and matched information sets, enabling ablation-style inference about incremental information content (Giacomini & White, 2006; Wang & Zhang, 2024). It also adopts a frozen-hyperparameter strategy for the rolling phase—hyperparameters are calibrated once and then held fixed—thereby precluding implicit test-time optimization and strengthening out-of-sample credibility (Chen & Guestrin, 2016; Tang et al., 2022).

Third, the paper operationalizes conformal inference as a governance tool for regime-aware risk monitoring and clarifies the boundary between information availability and usable predictive content in daily precious-metals return forecasting under strict rolling evaluation. By coupling transparent shock identification with sequential prediction intervals, the framework makes uncertainty escalation measurable and actionable for treasury and hedging routines, while the macro-augmentation ablation under conservative alignment documents when added complexity increases estimation risk rather than predictive value (Angelopoulos & Bates, 2021; Gibbs & Candès, 2024; Ying & Luo, 2025; Wang et al., 2026). The deterministic end-to-end pipeline—from data acquisition and feature construction to interval generation—also supports computational feasibility for near-real-time monitoring in industrial hedging and financial treasury settings (Tang et al., 2022; Luna et al., 2026).

Reference

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. Econometrica, 71(2), 579–625. https://doi.org/10.1111/1468-0262.00418

Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv. https://arxiv.org/abs/2107.07511

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

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Croushore, D. (2006). Forecasting with real-time macroeconomic data. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), Handbook of Economic Forecasting (Vol. 1, pp. 961–982). Elsevier. https://doi.org/10.1016/S1574-0706(05)01017-3

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. Journal of Business & Economic Statistics, 33(1), 1–9. https://doi.org/10.1080/07350015.2014.983236

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

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

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383–417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x

Foroni, C., Marcellino, M., & Schumacher, C. (2015). Unrestricted mixed data sampling (U-MIDAS): MIDAS regressions with unrestricted lag polynomials. Journal of Applied Econometrics, 30(3), 495–512. https://doi.org/10.1002/jae.2369

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. Journal of Econometrics, 131(1–2), 59–95. https://doi.org/10.1016/j.jeconom.2005.01.004

Gibbs, I., & Candès, E. J. (2021). Adaptive conformal inference under distribution shift. Advances in Neural Information Processing Systems. https://arxiv.org/abs/2106.00170

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. Econometrica, 74(6), 1545–1578. https://doi.org/10.1111/j.1468-0262.2006.00718.x

Hamilton, J. D., & Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. Journal of Econometrics, 64(1–2), 307–333. https://doi.org/10.1016/0304-4076(94)90067-1

Jabeur, S. B., Mefteh-Wali, S., & Viviani, J.-L. (2024). Forecasting gold price with the XGBoost algorithm and SHAP interaction values. Annals of Operations Research, 334(1–3), 679–699. https://doi.org/10.1007/s10479-021-04187-w

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523), 1094–1111. https://doi.org/10.1080/01621459.2017.1307116

Luna, M., Perez-Mon, O., & Becker, J. L. (2025). Forecasting and managing price volatility in salmon production: A hybrid system using conformal prediction and dynamic hedging. International Journal of Production Economics. https://doi.org/10.1016/j.ijpe.2025.109726

Naeem, M., Tiwari, A. K., Mubashra, S., & Shahbaz, M. (2019). Modeling volatility of precious metals markets by using regime-switching GARCH models. Resources Policy, 64, 101497. https://doi.org/10.1016/j.resourpol.2019.101497

Stark, T., & Croushore, D. (2002). Forecasting with a real-time data set for macroeconomists. Journal of Macroeconomics, 24(4), 507–531. https://doi.org/10.1016/S0164-0704(02)00062-9

Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world. Springer. https://doi.org/10.1007/b106715

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

Wang, X., Wu, J., & Liu, J. (2026). What drives precious metals pricing? An explainable mixed-frequency machine learning approach. Mineral Economics. https://doi.org/10.1007/s13563-025-00586-8

Xu, C., & Xie, Y. (2021). Conformal prediction interval for dynamic time-series. In Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139, pp. 11559–11569). https://proceedings.mlr.press/v139/xu21h.html

Ye, Y., Zhuang, X., Yi, C., Liu, D., & Tang, Z. (2025). Enhancing agricultural futures return prediction: Insights from rolling VMD, economic factors, and mixed ensembles. Agriculture, 15(11), 1127. https://doi.org/10.3390/agriculture15111127 (mdpi.com)

Ying, X., & Luo, B. (2025). Reducing forecast uncertainty in China's gold futures market through mixed-frequency volatility modeling. Finance Research Letters, 86, 108898. https://doi.org/10.1016/j.frl.2025.108898