AIMS Mathematics, 11(3): 5692–5718.
DOI: 10.3934/math.2026234
Received: 16 December 2025
Revised: 10 February 2026
Accepted: 27 February 2026
Published: 05 March 2026

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

Research article

Distribution-free uncertainty quantification for daily treasury yield curves with functional principal component forecasting and vector autoregression

Mervenur Sözen[1], **Fikriye Kabakcı**[2,*] **and Çağlar Sözen**[3]

[1] Independent researcher, Turkey; ORCID: 0000-0001-5603-5382

[2] Faculty of Arts and Sciences, Department of Mathematics, Recep Tayyip Erdoğan University, Rize, Turkey; ORCID: 0000-0001-6266-1902

[3] Görele School of Applied Sciences, Department of Finance and Banking, Giresun University, Giresun, Turkey; ORCID: 0000-0002-3732-5058

* **Correspondence:** Email: fikriye.kabakci@erdogan.edu.tr.

Abstract: We investigated one-step-ahead daily U.S. Treasury yield-curve forecasting and provided distribution-free uncertainty quantification for the entire term structure. Using constant-maturity yields from the Federal Reserve Bank of St. Louis (FRED), we first transformed the discrete maturity panel into a dense common maturity grid through a knot-consistent ridge-regularized cubic B-spline smoother, enabling coherent curve-level evaluation. For point prediction, we modeled the yield curve as a functional time series and forecasted functional principal component (FPCA) scores with a vector autoregression (VAR). We benchmarked FPCA–VAR against two widely used alternatives: The dynamic Nelson–Siegel (DNS) model and a raw-maturity PCA–VAR (RawPC–VAR) baseline. To quantify predictive uncertainty without imposing parametric distributional assumptions, we constructed rolling studentized conformal prediction bands using a simultaneous (sup-type) nonconformity score and a moving calibration window; the associated distribution-free validity was taken in the usual conformal (exchangeable) sense and treated as an operational benchmark—rather than a literal time-series guarantee—under temporal dependence. We therefore audited calibration directly on the test block and, to probe regime heterogeneity, implemented an ex-ante Mondrian conformal variant based on a curve-shock indicator that partitioned days into HIGH and LOW regimes. Out-of-sample results showed that FPCA–VAR achieved the lowest integrated squared error and yielded substantially tighter predictive bands than DNS, while Mondrian calibration improved interpretability by revealing and partially reducing regime-dependent coverage imbalances.

Keywords: yield-curve forecasting; functional principal component analysis; vector autoregression; conformal prediction bands; Mondrian conformal calibration
Mathematics Subject Classification: 62M10, 62G20

1. Introduction

The term structure of interest rates (the yield curve) is a central state variable in finance and macroeconomics. It underpins valuation, hedging, risk management, and monetary policy analysis, since it aggregates expectations about growth, inflation, and policy paths. From a statistical perspective, the yield curve is naturally a *functional object*: On each day t, observed yields define a curve $y_t(\tau)$ over maturity τ that exhibits strong cross-maturity coherence and evolves over time with business-cycle conditions, policy decisions, and time-varying risk premia. Despite the maturity of yield-curve modeling, *curve-wide* uncertainty quantification (UQ) for *daily* yield-curve forecasts remains comparatively less standardized. Workhorse approaches—dynamic Nelson–Siegel (DNS), affine no-arbitrage extensions, and broader affine term-structure models—typically deliver parametric forecast distributions, simulation-based intervals, or tenor-by-tenor bands that inherit structural and distributional assumptions. Existing daily empirical studies often report tenor-by-tenor intervals; in contrast, we target *simultaneous* maturity-wise control and evaluate uncertainty at the *curve level* on a common grid. At daily horizons, however, (i) one-step yield changes are small relative to noise, (ii) serial dependence is prominent, and (iii) heteroskedasticity and slow-moving structural change can be substantial. In such settings, it is valuable to complement model-based intervals with a transparent, assumption-lean uncertainty layer whose reliability can be diagnosed out-of-sample at the curve level.

This paper develops a unified and reproducible framework for one-step-ahead daily yield-curve forecasting together with curve-wide predictive bands on a common maturity domain. We first map each day's cross-section of observed maturities to a dense common grid via ridge-regularized B-spline smoothing. On this grid, we construct point forecasts using three baselines: (i) a functional principal component analysis (FPCA) pipeline with vector autoregression (VAR) score dynamics (FPCA–VAR), (ii) DNS, and (iii) a raw-maturity principal component analysis (PCA)–VAR baseline (RawPC–VAR), which applies PCA directly to the discretized maturity panel and forecasts the resulting principal component scores via VAR. The methodological focus is not to propose a new term-structure factorization, but to build a leak-safe pipeline that (a) places competing forecasters on the same grid-level functional target and (b) attaches a curve-wide distribution-free UQ layer that can be stress-tested under dependence and regime variation.

Building on each point forecaster, we construct rolling studentized conformal prediction bands for the entire curve. Conformal prediction calibrates predictive regions by inverting empirical quantiles of nonconformity scores and yields finite-sample marginal coverage under exchangeability [1–3]. Because classical exchangeability is violated in time series, we treat the usual conformal guarantee as an *exchangeable benchmark* rather than a literal time-series claim, and interpret nominal coverage as a practical target to be assessed by out-of-sample diagnostics. To improve practical stability under heteroskedasticity and drift, our implementation uses: (a) *rolling* calibration over a moving window to track local distributional change, and (b) *studentization* to normalize maturity-specific dispersion. Operationally, the band is driven by a studentized sup-type functional score (maximum over the maturity grid), which targets *simultaneous* (pathwise) control across maturities: Letting $e_t(\tau_g) = \tilde{y}_t(\tau_g) - \widehat{y}_{t|t-1}(\tau_g)$ denote the grid residual, the score is

$$A_t = \max_g \left\{ \frac{|e_t(\tau_g)|}{s_g(t)} \right\},$$

where $s_g(t)$ is a robust maturity-local scale (MAD) estimated from past residuals. Accordingly, we

report both curve-wise (pathwise) and maturity-wise (pointwise) coverage on the test block, together with average band width, to make calibration failures visible rather than implicit.

Yield dynamics can exhibit slow structural change and distributional drift. Rather than assuming strict stationarity or exchangeability, we implement the UQ layer in a strictly causal manner and use rolling calibration and studentization to stabilize the residual distribution locally. We therefore interpret nominal coverage as a practical target and assess it via test-block coverage diagnostics, complemented by dependence-robust inference for performance comparisons. In addition, we implement an *ex-ante Mondrian* (group-conditional) diagnostic [3,4]: Days are partitioned into *HIGH* and *LOW* curve-shock regimes using a shock proxy that is available at prediction time (computed from lagged realized curve changes). Mondrian calibration does not alter the point forecaster; it only selects the calibration pool for the band at time t. This yields a transparent conditional-calibration diagnostic: Pooled calibration may be overly conservative in calm periods yet exhibit undercoverage in turbulent periods, and regime-wise calibration can reduce such imbalance even when overall pathwise coverage remains challenging under dependence and finite buffers.

Finally, to compare point-forecast accuracy in a dependence-aware way, we evaluate models using daily integrated squared error (ISE) on the common grid and report (i) Diebold–Mariano tests with heteroskedasticity and autocorrelation consistent (HAC) long-run variance estimation, implemented via the Newey–West estimator, together with multiple-testing corrections, (ii) moving-block bootstrap confidence intervals for mean loss gaps, and (iii) model confidence set (MCS) elimination paths based on block-bootstrap resampling [5–9]. These procedures help ensure that conclusions about accuracy gaps are not artifacts of serial dependence or multiple-model uncertainty.

The manuscript's contributions are: (i) a reproducible, leak-safe pipeline for one-step-ahead daily yield-curve forecasting on a common maturity grid; (ii) rolling *studentized* conformal bands providing distribution-free, curve-wide UQ monitored via test-block calibration diagnostics; (iii) a transparent Mondrian (shock-conditional) variant that functions as a conditional-calibration diagnostic and pragmatic refinement; and (iv) dependence-robust inference for performance gaps via Diebold–Mariano (DM) tests with heteroskedasticity and autocorrelation consistent standard errors, moving-block bootstrap confidence intervals, and the MCS. The proposed perspective complements distribution-free band constructions for other financial curve objects (e.g., forward realized volatility paths) while addressing the yield curve's cross-sectional nature over maturity [10, 11].

Accordingly, we study one-step-ahead *daily* yield-curve forecasting together with *curve-wide* UQ under minimal distributional assumptions. Given a smoothed yield curve $y_t(\tau)$ on a common maturity grid, we produce a one-step-ahead point forecast $\hat{y}_{t|t-1}(\tau)$ and a data-driven band $\mathcal{B}_{t|t-1}(\tau)$ such that

$$\mathbb{P}(y_t(\cdot) \in \mathcal{B}_{t|t-1}(\cdot)) \approx 1 - \alpha,$$

where the approximation acknowledges temporal dependence and potential local distributional change. This leads to the following research questions:

RQ1: How do leading daily yield-curve forecasters (FPCA–VAR, RawPC–VAR, and DNS) compare in out-of-sample integrated accuracy on a common maturity grid?

RQ2: Can rolling studentized conformal bands deliver stable curve-wise (pathwise) coverage–width trade-offs at daily frequency under temporal dependence?

RQ3: Does an ex-ante curve-shock split reveal regime-conditional miscalibration under pooled conformal bands, and can Mondrian calibration reduce this regime imbalance?

RQ4: Which conclusions remain robust under temporal dependence when using DM tests with HAC standard errors, moving-block bootstrap confidence intervals (CIs), and MCS procedures?

2. Related literature

This section is organized to foreground the practical gap addressed by our design. While daily yield-curve forecasting is well studied, *curve-wide* (simultaneous over maturities) and *assumption-lean* uncertainty quantification at daily horizons remains comparatively less standardized. Many contributions rely on parametric simulation/likelihood assumptions or report tenor-by-tenor intervals without a unified grid-level curve target and explicit calibration evidence. Moreover, the time-series reality of yields—serial dependence with potentially local nonstationarity and regime variation—implies that nominal guarantees derived under exchangeability are not directly actionable. We therefore review (i) daily forecasters, (ii) functional representations and common-grid evaluation, (iii) modern conformal methods for dependent/shifted data and functional *band* constructions, and (iv) dependence-robust forecast comparison tools that support our robustness checks.

2.1. Yield-curve modeling and daily forecasting: Factor structure, dependence, and practical gaps

A large literature documents that a small number of factors account for most yield-curve variation. Early evidence in [12] emphasizes the empirical importance of level, slope, and curvature, motivating low-dimensional factor structures. The Nelson–Siegel representation [13] provides a parsimonious parametric cross-sectional form, later extended by [14]. The DNS model of [15] couples these factors with time-series dynamics (typically AR/VAR), yielding an interpretable decomposition with competitive forecasting performance. A related strand enforces no-arbitrage restrictions through affine term-structure models (ATSMs) with affine dynamics and risk premia; while structurally appealing, their out-of-sample performance can be sensitive to specification choices and risk-premia modeling [16]. Hybrid approaches retain tractability while imposing arbitrage-free structure; [17] develops an affine, arbitrage-free class of Nelson–Siegel models.

At daily frequency, one-step yield changes are often small relative to market microstructure noise and idiosyncratic shocks, while serial dependence and regime variation are prominent. Daily horizons therefore accentuate microstructure noise and regime-dependent volatility, motivating a calibration-audited UQ layer that can be assessed directly out-of-sample. Consequently, *curve-wide* UQ becomes a first-order complement to point accuracy: Bands should be interpretable over maturity and their calibration should be observable out-of-sample. This motivates our focus on a common-grid functional target and on calibration diagnostics for full-curve predictive statements, rather than proposing another parametric cross-sectional factorization.

2.2. Functional representations: FPCA, functional time series, and curve-level evaluation

Functional data analysis (FDA) formalizes curve-valued observations and provides tools to exploit smoothness and cross-domain dependence. Foundational references include [18] and, for dependent functional observations, [19]. FPCA decomposes $y_t(\tau)$ into a mean function plus a finite expansion

in eigenfunctions; the resulting score vectors can be forecast with standard time-series models. This "FPCA + time-series model" strategy is widely used in functional time-series forecasting [20, 21]. For yield curves, functional dynamic factor approaches estimate smooth loading curves and dynamic scores jointly; [22] provides a prominent application, and functional factor modeling continues to be developed for yield-curve-type panels, e.g., [23]. More recent work advances dynamic functional representations and forecasting methodology for dependent curves, e.g., [24, 25].

Because our empirical object is a *daily curve* evaluated on a dense maturity grid, we emphasize curve-level losses on a *common* maturity domain, summarized by integrated squared error (ISE) and its test-block average, mean integrated squared error (mISE). Because the grid is uniform, Euclidean PCA applied to grid vectors provides a close approximation to the L^2 inner product, while ISE is computed on the grid using trapezoidal integration weights. This ensures that point accuracy and uncertainty statements are assessed for the same functional target across FPCA–VAR, DNS, and RawPC–VAR under a single evaluation protocol.

2.3. Distribution-free uncertainty quantification: Conformal prediction and functional bands

Conformal prediction constructs predictive regions by calibrating nonconformity scores and inverting empirical quantiles [2,3]. In regression, split conformal and related variants provide marginal coverage without distributional assumptions under exchangeability [1]. For financial applications, the appeal is that calibration does not rely on Gaussianity, homoskedasticity, or a fully specified likelihood, which are routinely violated in interest-rate dynamics.

For functional targets, an additional desideratum is interpretability: Prediction sets should be *bands* over the domain rather than opaque set-valued objects. Recent work develops conformal band constructions for functional and multivariate functional objects, e.g., [26]. In our setting, a studentized subtype score over the maturity grid directly targets *simultaneous* (curve-wide) control and produces a visually interpretable band.

2.4. Conformal prediction under temporal dependence and distribution shift: Rolling calibration, studentization, and Mondrian diagnostics

Temporal dependence breaks classical exchangeability and can lead to systematic miscoverage if i.i.d. conformal is applied naively. A growing literature studies conformal prediction for time series and online settings by combining weak-dependence conditions with sequential/rolling designs and diagnostics, e.g., [27, 28]. Beyond dependence, nonstationarity and distribution shift are central stylized features in macro-finance; adaptive calibration mechanisms explicitly designed for distribution shift motivate rolling/online variants, e.g., [29, 30]. Rolling windows act as a simple adaptive mechanism under gradual drift, trading strict guarantees for practical stability in time-series environments. These insights motivate our emphasis on strict time ordering, rolling/sliding-window calibration to track gradual drift, and studentized scores to stabilize widths under maturity-dependent and time-varying volatility.

To diagnose and potentially mitigate regime-dependent miscoverage, we employ *Mondrian conformal prediction*, where calibration is performed within groups defined by observable covariates [3, 4]. In yield-curve applications, natural grouping variables include level/slope shocks or volatility proxies; our ex-ante HIGH/LOW curve-shock partition provides a transparent conditional-

calibration diagnostic for turbulent vs. calm periods.

2.5. Inference for dependent forecast evaluation and model comparison

Evaluating yield-curve forecasters calls for inference tools robust to serial dependence. The Diebold–Mariano test [5] provides a standard approach for testing equal predictive accuracy, and dependent implementations typically rely on HAC long-run variance estimation such as Newey–West [9]. Resampling methods quantify uncertainty in dependent forecast evaluation. Moving-block bootstrap procedures preserve local dependence by resampling contiguous blocks [7, 8]. Finally, when multiple models compete, multiple-comparison uncertainty becomes central. The MCS of [6] identifies a set of superior models at a chosen confidence level and is well-suited to macro-finance forecast comparisons where performance differences can be small relative to sampling variability.

In line with this literature, we treat exchangeability-based conformal validity as a benchmark reference and adopt rolling/studentized calibration as a practical device for dependent and slowly drifting time-series environments, while auditing calibration explicitly via test-block coverage diagnostics [31, 32].

3. Data and preprocessing

3.1. Data source and sample construction

We use U.S. Treasury constant-maturity yields from the Federal reserve economic data (FRED) database maintained by the Federal Reserve Bank of St. Louis [33]. The raw series are reported in annualized percent units; throughout, we convert them to decimal yields by dividing by 100. The maturity cross-section contains $P = 11$ tenors spanning one month to thirty years,

$$m_j \in \left\{\tfrac{1}{12}, \tfrac{3}{12}, \tfrac{6}{12}, 1, 2, 3, 5, 7, 10, 20, 30\right\} \text{ (years)},$$

corresponding to the FRED identifiers `DGS1MO, DGS3MO, DGS6MO, DGS1, DGS2, DGS3, DGS5, DGS7, DGS10, DGS20, DGS30`.

Let $\{t_i\}_{i=1}^{T}$ denote the business-day index in the merged panel. Because some dates contain missing values for at least one maturity, we form a balanced panel by complete-case alignment. We merge all $P = 11$ series and retain only those dates for which the full cross-section is simultaneously observed. The resulting aligned sample spans 2002–01–02 to 2024–12–31 (inclusive) and contains $T = 5754$ daily yield curves.

Following standard practice for dependent time series, we use a strictly chronological split into training, calibration, and test blocks. In our aligned sample, the exact date ranges are: train = 2002–01–02 to 2015–10–15, calibration = 2015–10–16 to 2020–05–26, and test = 2020–05–27 to 2024–12–31 (see Table 1). The corresponding block sizes are 3452/1151/1151 days (approximately a 60%/20%/20% split). The calibration block initializes the rolling conformal residual buffers in Section 4.4, and all model estimations and evaluations are performed strictly forward in time (no look-ahead).

Table 1. Data summary for the yield-curve sample (daily FRED constant-maturity Treasury series).

Item	Value	Notes	Item	Value	Notes
Data source	FRED	Federal Reserve Economic Data [33].	Maturities (P)	11	Tenors spanning one month to thirty years.
Series identifiers	`DGS1MO`, `DGS3MO`, `DGS6MO`, `DGS1`, `DGS2`, `DGS3`, `DGS5`, `DGS7`, `DGS10`, `DGS20`, `DGS30`	Constant-maturity U.S. Treasury yields.	Maturity range	1M–30Y	$m_j \in \{1/12, \ldots, 30\}$ years.
Units	Percent → decimal	Divide all yields by 100 before modeling.	Frequency	Daily	Business days (FRED calendar).
Sample start	2002–01–02	First aligned day with complete cross-section.	Train window	2002–01–02 to 2015–10–15	Chronological block (exact dates).
Sample end	2024–12–31	End date fixed.	Calibration window	2015–10–16 to 2020–05–26	Chronological block; initializes rolling conformal buffers.
Sample size (days)	5754	Daily curves after complete-case alignment.	Test window	2020–05–27 to 2024–12–31	Chronological block; strictly out-of-sample evaluation.
Split sizes	3452 / 1151 / 1151	Train / calibration / test (approx. 60%/20%/20%).	Common grid (G)	121	Uniform maturity grid on [1/12, 30] years.

3.2. Common functional grid via ridge-regularized B-spline smoothing

The observed yield curve on each day is available only on the discrete maturity set $\{m_j\}_{j=1}^{P}$. To (i) compare all forecasters on an identical maturity domain and (ii) compute integrated losses and curve-wide conformal bands consistently, we map each daily cross-section to a dense common grid $\{\tau_g\}_{g=1}^{G}$ with $G = 121$ points, uniformly spanning $[1/12, 30]$ years:

$$\tau_g = \tfrac{1}{12} + (g-1)\frac{30 - \frac{1}{12}}{G-1}, \qquad g = 1, \ldots, G.$$

For each day t, let $\boldsymbol{y}_t = (y_t(m_1), \ldots, y_t(m_P))^\top \in \mathbb{R}^P$ be the observed maturity vector (in decimals). We construct a cubic B-spline basis evaluated at the observed maturities, $B \in \mathbb{R}^{P\times d}$, using a fixed knot/boundary specification shared across all dates; in the empirical pipeline we use $d = 8$ basis functions (`df=8`, with intercept). To stabilize estimation given only $P = 11$ observed maturities, spline coefficients are estimated by ridge-regularized least squares:

$$\widehat{\boldsymbol{c}}_t = \arg\min_{\boldsymbol{c}\in\mathbb{R}^d} \|\boldsymbol{y}_t - B\boldsymbol{c}\|_2^2 + \lambda_{\text{ridge}}\|\boldsymbol{c}\|_2^2 = (B^\top B + \lambda_{\text{ridge}} I_d)^{-1} B^\top \boldsymbol{y}_t, \qquad \lambda_{\text{ridge}} = 10^{-4}. \tag{3.1}$$

The smoothed curve on the common grid is obtained by evaluating the *same* spline basis (same knots and boundary knots) on the grid. Let $B_g \in \mathbb{R}^{G\times d}$ denote the same B-spline basis as B, evaluated at the common grid points $\{\tau_g\}_{g=1}^G$.

$$\tilde{\boldsymbol{y}}_t^g = (\tilde{y}_t(\tau_1), \ldots, \tilde{y}_t(\tau_G))^\top = B_g \widehat{\boldsymbol{c}}_t \in \mathbb{R}^G. \tag{3.2}$$

Because the knot/boundary specification depends only on the maturity axis and is fixed ex ante, the smoothing step uses no future yield information (no look-ahead). Reusing identical knot/boundary specifications in B and B_g ensures a stable observed-to-grid mapping over time and guarantees that all methods are evaluated on exactly the same maturity domain.

For integrated loss computations on the common grid, we use trapezoidal maturity weights (equivalently, grid-spacing/Riemann-type weights on a uniform grid), so that ISE approximates the maturity integral of squared forecast error on $[1/12, 30]$. The resulting functional panel $\{\tilde{\boldsymbol{y}}_t^g\}_{t=1}^T$ is the common representation used throughout the paper: FPCA–VAR operates directly on these grid curves; DNS is estimated on observed maturities but evaluated on the grid; and RawPC–VAR forecasts the maturity panel and then applies the same mapping. All loss computations (ISE) and conformal bands are constructed and assessed on the common grid.

We denote the smoothed yield curve on the common grid by $\tilde{\boldsymbol{y}}_t^g = (\tilde{y}_t(\tau_1), \ldots, \tilde{y}_t(\tau_G))^\top$. At prediction date t, using information available up to $t-1$, its one-step-ahead point forecast is denoted by $\widehat{\boldsymbol{y}}_{t|t-1}^g = (\widehat{y}_{t|t-1}(\tau_1), \ldots, \widehat{y}_{t|t-1}(\tau_G))^\top$. For notational brevity, we may occasionally write $\widehat{\boldsymbol{y}}_t^g$ as shorthand for $\widehat{\boldsymbol{y}}_{t|t-1}^g$ when the one-step-ahead conditioning is clear from context. The grid-level residual is defined by $e_t(\tau_g) = \tilde{y}_t(\tau_g) - \widehat{y}_{t|t-1}(\tau_g)$ for $g = 1, \ldots, G$.

4. Methods

4.1. Problem setup, preprocessing, and notation

Let $\{y_t(m_j)\}_{j=1}^P$ denote the observed daily U.S. Treasury yields (in decimals) at maturities $\{m_j\}_{j=1}^P \subset [1/12, 30]$ (in years), with $P = 11$ in our application. Following the spline mapping step in Section 3.2, each day t is represented by a smoothed yield curve evaluated on a common dense maturity grid $\{\tau_g\}_{g=1}^G$ with $G = 121$:

$$\tilde{\boldsymbol{y}}_t^g := (\tilde{y}_t(\tau_g))_{g=1}^G \in \mathbb{R}^G, \qquad \tau_g \in [1/12, 30].$$

Our task is one-step-ahead curve prediction: At prediction date t, using information available up to $t-1$, we output (i) a point forecast $\widehat{\boldsymbol{y}}_{t|t-1}^g$ for day t and (ii) a simultaneous predictive band $\widehat{\mathcal{B}}_{t|t-1} \subset \mathbb{R}^G$ for the entire curve.

We emphasize that $\tilde{\boldsymbol{y}}_t^g$ is the common grid-level *target* obtained by spline smoothing of the observed cross-section, whereas $\widehat{\boldsymbol{y}}_{t|t-1}^g$ is the corresponding one-step-ahead *forecast* on the same grid. To keep notation light, we will often write $\widehat{\boldsymbol{y}}_t^g$ as shorthand for $\widehat{\boldsymbol{y}}_{t|t-1}^g$ when the one-step-ahead conditioning is clear from context. Define the grid residual function by

$$e_t(\tau_g) = \tilde{y}_t(\tau_g) - \widehat{y}_{t|t-1}(\tau_g), \qquad g = 1, \ldots, G,$$

and the residual vector $\boldsymbol{e}_t = (e_t(\tau_g))_{g=1}^G \in \mathbb{R}^G$. Throughout, $\|\boldsymbol{v}\|_\infty := \max_{1\le g\le G} |v_g|$ denotes the sup-norm on the grid, and MAD$(\cdot)$ denotes the median absolute deviation.

We use a strictly chronological split into training, calibration, and test blocks. All tuning choices are fixed throughout the empirical pipeline. Specifically, we target a proportion of variance explained (PVE) of 0.995, use a maximum retained dimension of $K_{\max} = 12$, impose a VAR lag cap of $p_{\max} = 5$, and refit every 20 trading days. For uncertainty quantification, we set the miscoverage level to $\alpha = 0.10$, the rolling calibration window to $W = 750$, the studentization offset to $\varepsilon = 10^{-6}$, and the Mondrian high-shock threshold to the empirical quantile $q_{\text{high}} = 0.80$. The rolling-window length $W = 750$ is selected based on the sensitivity analysis reported in Table 6, balancing curve-wise coverage against average band width.

To avoid any "outside-the-framework" interpretation, we emphasize a clear separation between (a) the *forecasting + UQ layer* and (b) the *evaluation/inference layer*. The forecasting component consists of the point predictor (FPCA–VAR / DNS / RawPC–VAR) together with the rolling, studentized conformal band construction and the ex-ante Mondrian (HIGH/LOW) split, which together define a strictly causal, distribution-free *uncertainty quantification* procedure for the entire curve. In contrast, the Diebold–Mariano tests with HAC standard errors, moving-block bootstrap confidence intervals, and the MCS are *standard dependence-robust tools* used solely to quantify uncertainty in *comparative forecast evaluation* under serial dependence. These inferential procedures do not modify the forecasting rules; they provide an auditable, dependence-aware assessment of whether observed performance gaps are statistically meaningful.

4.2. Point forecasting models

We compare three standard one-step curve forecasters: (i) a functional FPCA pipeline with multivariate score dynamics (FPCA–VAR), (ii) a DNS, and (iii) a panel PCA baseline applied to the raw maturity vector (RawPC–VAR). All models are initialized on the training block and evaluated on the common grid.

Although each model is initialized on the training block, out-of-sample forecasting is implemented in a strictly chronological, real-time fashion: The forecast for day t uses only information available up to $t - 1$. To balance computational cost with adaptivity, time-series parameters (such as the VAR dynamics for the FPCA, DNS, and RawPC score vectors) are re-estimated every 20 trading days on an expanding information set; forecasts between re-estimation dates rely on the most recently fitted specification. To prevent implicit look-ahead, all cross-sectional objects are fixed ex ante: The FPCA loading functions and mean curve are estimated once on the training block and then held fixed, the RawPC loading vectors and maturity-level mean are likewise estimated on the training block and held fixed, and the DNS decay parameter λ is selected on the training block and kept fixed throughout the evaluation period.

4.2.1. FPCA–VAR forecasting (functional baseline)

Let $\boldsymbol{\mu} \in \mathbb{R}^G$ be the pointwise mean of the training curves on the common grid. The FPCA basis $\Phi = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K) \in \mathbb{R}^{G \times K}$ is computed from the centered training matrix via singular value decomposition (SVD), and the smallest K is selected such that the cumulative proportion of variance explained (PVE) exceeds 0.995, subject to $K \leq K_{\max} = 12$. Define the FPCA score vector for day t by

$$\boldsymbol{s}_t = \Phi^\top(\tilde{\boldsymbol{y}}_t^g - \boldsymbol{\mu}) \in \mathbb{R}^K. \tag{4.1}$$

We model $\{\boldsymbol{s}_t\}$ using a VAR(p) with intercept. The lag order p is selected on the training block by the Akaike information criterion (AIC), subject to the cap $p_{\max} = 5$, and the VAR is refit every 20 trading days on the expanding score history. Let $\widehat{\boldsymbol{s}}_{t|t-1}$ be the one-step-ahead score forecast. The curve forecast is reconstructed as

$$\widehat{\boldsymbol{y}}^{g}_{t|t-1} = \boldsymbol{\mu} + \Phi \widehat{\boldsymbol{s}}_{t|t-1}. \tag{4.2}$$

4.2.2. DNS baseline

We implement a DNS representation on the observed maturities. Let $\tilde{m}_j = 12m_j$ denote month-scale maturities. For decay $\lambda > 0$ (units: month^{-1}), define the Nelson–Siegel loadings

$$\boldsymbol{b}(\tilde{m}; \lambda) = \left(1,\ \frac{1 - e^{-\lambda\tilde{m}}}{\lambda\tilde{m}},\ \frac{1 - e^{-\lambda\tilde{m}}}{\lambda\tilde{m}} - e^{-\lambda\tilde{m}}\right). \tag{4.3}$$

For each day t and fixed λ, the factor vector $\boldsymbol{\beta}_t = (\beta_{0,t}, \beta_{1,t}, \beta_{2,t})^\top$ is estimated by ordinary least squares (OLS) on the P observed maturities. We select λ by grid search, minimizing the average training cross-sectional sum of squared errors (SSE) over

$$\Lambda = \{0.005, 0.010, \ldots, 0.250\} \text{ (month}^{-1}\text{)}.$$

We then forecast $\boldsymbol{\beta}_t$ with a VAR(p) (AIC-selected, $p \le p_{\max} = 5$, refit every 20 days on the expanding factor history). Let $\widehat{\boldsymbol{\beta}}_{t|t-1}$ denote the one-step-ahead factor forecast. Finally, we map $\widehat{\boldsymbol{\beta}}_{t|t-1}$ to the common grid by evaluating (4.3) at $\tilde{\tau}_g = 12\tau_g$ (months) for $g = 1, \ldots, G$, yielding the grid-level curve forecast $\widehat{\boldsymbol{y}}^{g}_{t|t-1}$.

4.2.3. RawPC–VAR baseline

We apply PCA directly to the P-dimensional raw maturity panel on the training block, retain the smallest number of components achieving PVE ≥ 0.995 (cap $K_{\max} = 12$), and forecast the resulting PC scores via a VAR(p) with AIC selection and refitting every 20 days (again $p \le p_{\max} = 5$) on an expanding score history. The maturity-level one-step-ahead forecast is then mapped to the common grid using the same spline smoother as in preprocessing (Section 3.2), enabling like-for-like grid-level evaluation.

4.3. Loss function and evaluation targets

Accuracy comparisons are performed on the common grid using the daily ISE:

$$\mathrm{ISE}_t = \sum_{g=1}^{G} w_g \Big(\tilde{y}_t(\tau_g) - \widehat{y}_{t|t-1}(\tau_g)\Big)^2, \tag{4.4}$$

where w_g are trapezoidal maturity weights on the uniform grid. With spacing $\Delta\tau = (30 - \frac{1}{12})/(G-1)$, we use

$$w_1 = w_G = \Delta\tau/2, \qquad w_g = \Delta\tau \ (g = 2, \ldots, G-1),$$

so that ISE_t approximates $\int_{1/12}^{30} (\tilde{y}_t(\tau) - \widehat{y}_{t|t-1}(\tau))^2 \, d\tau$.

4.4. Rolling studentized conformal bands (pooled)

Given a point forecaster producing $\widehat{\boldsymbol{y}}^{g}_{t|t-1}$, define grid residuals $\boldsymbol{e}_t = \tilde{\boldsymbol{y}}^{g}_{t} - \widehat{\boldsymbol{y}}^{g}_{t|t-1} \in \mathbb{R}^G$. At each test day t, we maintain a rolling calibration buffer of length $W = 750$,

$$\mathcal{E}_t = \{\boldsymbol{e}_{t'} : \ t - W \le t' \le t - 1\}, \qquad n = |\mathcal{E}_t|. \tag{4.5}$$

The buffer is initialized from the calibration block. We target miscoverage $\alpha = 0.10$.

We estimate a robust maturity-local scale using MAD:

$$s_g(t) = 1.4826\,\mathrm{MAD}(\{e_{t'}(\tau_g) : \boldsymbol{e}_{t'} \in \mathcal{E}_t\}) + \varepsilon, \qquad \varepsilon = 10^{-6}. \tag{4.6}$$

Here, $\mathrm{MAD}(x) := \mathrm{median}(|x - \mathrm{median}(x)|)$ and 1.4826 is the conventional Gaussian-consistency normalization. Because $s_g(t)$ is computed using residuals up to $t - 1$ only, the construction is strictly causal (no look-ahead).

For prediction time t, define the studentized subtype score for any residual curve $\boldsymbol{e} \in \mathbb{R}^G$ as

$$A_t(\boldsymbol{e}) = \max_{1\le g\le G} \frac{|e(\tau_g)|}{s_g(t)}. \tag{4.7}$$

We compute calibration scores $\{A_t(\boldsymbol{e}_{t'})\}_{\boldsymbol{e}_{t'}\in\mathcal{E}_t}$ and set

$$k = \lceil (n+1)(1-\alpha) \rceil, \qquad A_{(1)} \le \cdots \le A_{(n)} \text{ the sorted calibration scores } \{A_t(\boldsymbol{e}_{t'})\}.$$

Set the finite-sample corrected threshold as the k-th order statistic:

$$q_t = A_{(k)}. \tag{4.8}$$

The pooled rolling studentized band for day t is

$$\widehat{\mathcal{B}}^{\mathrm{pool}}_{t|t-1} = \Big\{\boldsymbol{y} \in \mathbb{R}^G : \ |y(\tau_g) - \widehat{y}_{t|t-1}(\tau_g)| \le q_t\, s_g(t), \ \forall g\Big\}. \tag{4.9}$$

Let $\mathcal{T}_{\mathrm{te}}$ denote the test index set with size T_{te}. We report $\mathrm{mISE} := T_{\mathrm{te}}^{-1} \sum_{t\in\mathcal{T}_{\mathrm{te}}} \mathrm{ISE}_t$. For a band with grid-wise lower/upper limits $\{(L_t(\tau_g), U_t(\tau_g))\}_{g=1}^{G}$, we define

$$\mathrm{PathCov} := \frac{1}{T_{\mathrm{te}}} \sum_{t\in\mathcal{T}_{\mathrm{te}}} \mathbf{1}\Big\{\forall g : \ L_t(\tau_g) \le \tilde{y}_t(\tau_g) \le U_t(\tau_g)\Big\}, \tag{4.10}$$

$$\mathrm{PtCov} := \frac{1}{G} \sum_{g=1}^{G} \left(\frac{1}{T_{\mathrm{te}}} \sum_{t\in\mathcal{T}_{\mathrm{te}}} \mathbf{1}\{L_t(\tau_g) \le \tilde{y}_t(\tau_g) \le U_t(\tau_g)\} \right), \tag{4.11}$$

$$\mathrm{AvgBW} := \frac{1}{G} \sum_{g=1}^{G} \left(\frac{1}{T_{\mathrm{te}}} \sum_{t\in\mathcal{T}_{\mathrm{te}}} (U_t(\tau_g) - L_t(\tau_g)) \right). \tag{4.12}$$

For studentized rolling conformal bands, $U_t(\tau_g) - L_t(\tau_g) = 2\, q_t\, s_g(t)$, and we also report $\bar{q} := T_{\mathrm{te}}^{-1} \sum_{t\in\mathcal{T}_{\mathrm{te}}} q_t$.

4.5. Exchangeable conformal benchmark guarantee

We state the standard distribution-free guarantee under exchangeability as a theoretical baseline (and *not* as a literal time-series claim).

Assumption 4.1 (Exchangeability of studentized scores (benchmark))**.** *For a given prediction time t, the scores $\{A_t(\boldsymbol{e}_{t-W}), \ldots, A_t(\boldsymbol{e}_{t-1}), A_t(\boldsymbol{e}_t)\}$ are exchangeable conditional on the fitted point forecaster (trained using data up to $t-1$).*

We identify a grid vector $\boldsymbol{y} \in \mathbb{R}^G$ with its components $y(\tau_g)$ on the grid.

Proposition 4.1 (Exchangeable conformal benchmark: Functional band coverage)**.** *Under Assumption 4.1, the band in* (4.9) *satisfies*

$$\mathbb{P}\Big(\tilde{\boldsymbol{y}}_t^g \in \widehat{\mathcal{B}}_{t|t-1}^{\mathrm{pool}}\Big) \geq 1-\alpha.$$

Proof. Define $A_t^{\mathrm{new}} := A_t(\boldsymbol{e}_t)$ and let q_t be defined as in (4.8). The event $\tilde{\boldsymbol{y}}_t^g \notin \widehat{\mathcal{B}}_{t|t-1}^{\mathrm{pool}}$ is equivalent to $A_t^{\mathrm{new}} > q_t$. Under exchangeability, the rank of A_t^{new} among the $n+1$ scores $\{A_t(\boldsymbol{e}_{t-W}), \ldots, A_t(\boldsymbol{e}_{t-1}), A_t^{\mathrm{new}}\}$ is uniform on $\{1, \ldots, n+1\}$, and the standard split conformal argument yields $\mathbb{P}(A_t^{\mathrm{new}} > q_t) \leq \alpha$. □

4.6. Approximate validity under weak dependence and local nonstationarity

Daily yields are temporally dependent and may exhibit slow-moving structural change, so strict exchangeability is not literally satisfied. Our theoretical stance is therefore deliberately conservative: Proposition 4.1 is treated as an *exchangeable benchmark* (a reference point), not as a literal time-series guarantee. In dependent or non-exchangeable settings, recent conformal theory provides conditions under which conformal calibration can remain approximately valid, typically incurring a (possibly small) coverage deviation that depends on the strength of dependence or departure from exchangeability; see, e.g., [31] and the time-series conformal literature [27, 28, 32].

Operationally, our design follows two widely used stabilizing principles for time series UQ. First, we use a *rolling/sliding calibration window* to adapt to gradual distributional drift, which is aligned with the general rationale of adaptive/online conformal calibration under distribution shift [29, 30]. Second, we apply *studentization* via robust maturity-local scales to mitigate maturity-dependent and time-varying dispersion. Together, rolling calibration and studentization aim to make the empirical distribution of recent nonconformity scores locally stable over moderate windows, consistent with a *locally stationary* interpretation of daily yield dynamics [34].

Remark 4.1. Because exact finite-sample conformal validity is guaranteed under exchangeability but not under general dependence, we interpret nominal coverage as an operational target and evaluate it via explicit test-block diagnostics. Moreover, our MAD-based studentization estimates $s_g(t)$ from the rolling calibration window; thus, even under i.i.d. data, exact finite-sample conformal validity would generally require a fully symmetric (transductive) normalization, further reinforcing our benchmark–audit interpretation. Accordingly, our empirical results report both curve-wise (pathwise) and maturity-wise (pointwise) coverage, together with band widths and the time variation of rolling thresholds, so that any miscalibration becomes directly observable rather than implicit. Dependence-robust inference for point-forecast comparisons is provided separately via HAC and block-resampling tools in Section 4.8.

4.7. Mondrian conformal with ex-ante curve-shock regimes

To diagnose regime heterogeneity and improve *conditional* calibration, we use a transparent curve-shock proxy that is computable at prediction time and affects *only* the conformal calibration pool (not the point forecaster).

Let $\{\tau_g\}_{g=1}^{G}$ be the common maturity grid and define the (realized) one-day curve change $\Delta\tilde{y}_t(\tau_g) = \tilde{y}_t(\tau_g) - \tilde{y}_{t-1}(\tau_g)$. We summarize the magnitude of this shock by

$$S_t \;=\; \max_{1\le g\le G} \big|\Delta\tilde{y}_t(\tau_g)\big|. \tag{4.13}$$

Let c be the empirical q_{high}-quantile of $\{S_t\}$ computed over the combined training and calibration (TRAIN+CAL) period (in our application $q_{\text{high}} = 0.80$). The regime label used for forecasting day t is then defined *without look-ahead* as

$$R_t \;=\; \begin{cases} \text{HIGH}, & S_{t-1} > c, \\ \text{LOW}, & S_{t-1} \le c, \end{cases} \tag{4.14}$$

so R_t is known at time t using only information up to $t-1$. Here, $r \in \{\text{LOW}, \text{HIGH}\}$ indexes the *shock-conditional calibration group* used to compute $(q_t^r, s_g^r(t))$; the underlying point forecast $\widehat{\boldsymbol{y}}_{t|t-1}^{\,g}$ is unchanged.

We maintain two regime-specific residual buffers, $\mathcal{E}_t^{\text{LOW}}$ and $\mathcal{E}_t^{\text{HIGH}}$. Importantly, the window length $W = 750$ is applied *within regime*: Each buffer stores the most recent W residual curves *from that regime* (not necessarily the last W calendar days), so the effective calendar span used for calibration can exceed W during persistent LOW/HIGH spells. Within regime $r \in \{\text{LOW}, \text{HIGH}\}$ we compute maturity-local MAD scales

$$s_g^r(t) = 1.4826\,\text{MAD}\Big(\{e_{t'}(\tau_g) : \boldsymbol{e}_{t'} \in \mathcal{E}_t^r\}\Big) + \varepsilon, \qquad \varepsilon = 10^{-6},$$

and define the regime-wise nonconformity score (studentized subtype)

$$A_t^r(\boldsymbol{e}) = \max_{1\le g\le G} \frac{|e(\tau_g)|}{s_g^r(t)}. \tag{4.15}$$

Using only $\mathcal{E}_t^r$, we compute calibration scores $\{A_t^r(\boldsymbol{e}_{t'})\}_{\boldsymbol{e}_{t'}\in\mathcal{E}_t^r}$ and set the regime-specific threshold q_t^r via the same k-of-n rule as in (4.8). The Mondrian band for day t is then

$$\widehat{\mathcal{B}}_{t|t-1}^{\text{mond}} = \Big\{\boldsymbol{y} \in \mathbb{R}^G :\; |y(\tau_g) - \widehat{y}_{t|t-1}(\tau_g)| \le q_t^{R_t}\, s_g^{R_t}(t),\; \forall g\Big\}. \tag{4.16}$$

This design stabilizes quantile estimation within each regime and avoids mixing residual distributions across regimes.

Let $n_t^r := |\mathcal{E}_t^r|$ denote the size of the active regime buffer. If $n_t^{R_t} < n_{\min}$, we revert to pooled calibration for that date to preserve numerical stability; in our implementation, we set $n_{\min} = 30$. Otherwise, we use the regime-specific $(q_t^{R_t}, s_g^{R_t}(t))$ as in (4.16).

Band widths are driven by the distribution of *extreme* studentized residuals through the max-score and the resulting q_t^r, and also by the maturity-local dispersion $s_g^r(t)$. Hence, two forecasters can have similar mean ISE yet materially different average band widths if one exhibits (i) larger tail behavior in $\max_g |e(\tau_g)|/s_g^r(t)$ (inflating q_t^r) and/or (ii) larger residual dispersion across maturities (inflating $s_g^r(t)$). We therefore report accuracy together with $(\bar{q}, \text{AvgBW})$ and coverage to disentangle these mechanisms.

4.8. Statistical comparison: DM tests, block bootstrap CIs, and MCS

We compare point-forecast accuracy via daily ISE losses and Diebold–Mariano tests on loss differentials. To account for serial dependence, we estimate the long-run variance with a Newey–West HAC estimator using lag $\lfloor T_{\text{te}}^{1/3} \rfloor$ on the test block. Multiple-comparison adjustments are implemented using the Holm family-wise error rate (FWER) procedure and Benjamini–Hochberg false discovery rate (BH/FDR) control.

Uncertainty in mean loss gaps is quantified by moving-block bootstrap confidence intervals: We resample overlapping blocks of length B to form pseudo-series of length T_{te}, preserving local dependence. In the empirical pipeline we use $R = 2000$ replications and report a sensitivity grid of block lengths. We additionally apply the MCS procedure using block-bootstrap resampling and the $T_{\max}$ statistic, iteratively eliminating inferior models until equal predictive ability cannot be rejected at level $\alpha_{\text{MCS}} = 0.10$.

The use of *block* resampling is motivated by the serial dependence of daily loss differentials: i.i.d. resampling would break temporal dependence and can severely understate uncertainty. Moving-block and related block bootstrap schemes are classical dependence-preserving devices and are well known to yield consistent inference for sample means (and mean loss gaps) under weak dependence / mixing-type conditions in stationary time series; see [7] and the monograph treatment in [8]. As a robustness check on resampling design, we report a sensitivity grid over block lengths; alternative dependence-preserving variants (e.g., stationary bootstrap) are also available in the literature [35], but our moving-block implementation is sufficient for the dependence-audited comparisons reported here.

5. Empirical results

This section reports the strictly out-of-sample evaluation on the test (TEST) block covering 2020–05–27 to 2024–12–31 (Table 1) and addresses the research questions posed in the Introduction. All accuracy and uncertainty results are computed on the common maturity grid described in Section 3.2, using the ISE loss (4.4) and the pooled/Mondrian rolling conformal bands (4.9)–(4.16). Our emphasis is on (i) grid-level curve accuracy, (ii) curve-wide (simultaneous) distribution-free uncertainty quantification, and (iii) dependence-aware robustness checks via HAC Diebold–Mariano tests, moving-block bootstrap, and the MCS (Section 4.8).

5.1. Main out-of-sample accuracy and pooled conformal performance (RQ1–RQ2)

Table 2 reports out-of-sample mISE for point forecasts, together with pooled rolling studentized conformal performance (curve-wise/pathwise coverage, pointwise coverage, average band width, and the average rolling conformal threshold). Across all criteria, FPCA–VAR is the top performer. It attains the lowest mISE (1.28766×10^{-4}), improving upon RawPC–VAR by about 3.6%, while DNS is substantially worse (9.59611×10^{-4}), i.e., roughly 7.5× larger than FPCA–VAR.

Pooled rolling bands are close to nominal in the curve-wise sense for FPCA and RawPC (pathwise coverage 0.883 and 0.880 vs. target 0.90), while DNS exhibits lower curve-wise coverage (0.798) despite being substantially wider on average. Pointwise coverage is conservative for all models (around 0.98), consistent with the use of a studentized subtype (simultaneous) nonconformity score, which targets uniform control over maturities rather than marginal coverage at a single tenor.

Table 2. Out-of-sample accuracy and pooled rolling studentized conformal performance (TEST).

Estimator	mISE	PathCov	PtCov	AvgBW	$\bar{q}$
FPCA–VAR	0.000128766	0.883	0.982	0.0126032	5.5650
RawPC–VAR	0.000133566	0.880	0.981	0.0132164	5.6393
DNS	0.000959611	0.798	0.987	0.0382960	5.5530

Empirical answers to the research questions (TEST block).

RQ1 Point accuracy: FPCA–VAR achieves the lowest mISE on the common grid, with RawPC–VAR second and DNS substantially worse.

RQ2 Distribution-free UQ (pooled): Rolling studentized conformal bands provide stable curve-wide uncertainty; pointwise coverage is conservative due to simultaneous calibration.

RQ3 Regime heterogeneity: Pooled calibration masks LOW/HIGH differences; Mondrian calibration exposes (and partially mitigates) conditional miscalibration.

RQ4 Robustness under dependence: HAC DM tests, block-bootstrap CIs, and MCS jointly support FPCA–VAR as the top performer under ISE loss.

5.2. Mondrian conformal bands and conditional diagnostics

Table 3 summarizes the rolling Mondrian (group-conditional) conformal variant based on an *ex-ante* curve-shock regime split (Section 4.7). Relative to pooled calibration, the Mondrian scheme yields slightly narrower average band width (AvgBW) for FPCA and RawPC (e.g., FPCA 0.01260 → 0.01135), while maintaining conservative pointwise coverage.

Table 3. Out-of-sample accuracy and rolling studentized Mondrian conformal performance (TEST).

Estimator	**mISE**	**PathCov**	**PtCov**	**AvgBW**	$\bar{q}$
FPCA–VAR	**0.000128766**	0.8510	0.9760	**0.0113532**	**4.9167**
RawPC–VAR	0.0001	**0.858**	**0.978**	0.0119	4.9644
DNS	0.0010	0.7730	0.9770	0.0360	5.0561

Its main value is *conditional calibration*. Table 10 shows that pooled bands can be materially more conservative in LOW-shock periods while under-covering in HIGH-shock periods for FPCA and RawPC (e.g., FPCA PathCov 0.918 in LOW vs. 0.787 in HIGH). Mondrian calibration reduces this regime imbalance (Table 11) by calibrating within regime, yielding more even curve-wise coverage across LOW and HIGH conditions (FPCA: 0.841 vs. 0.877; RawPC: 0.845 vs. 0.890). At the same time, overall curve-wise coverage under Mondrian can fall further below nominal under temporal dependence and finite within-regime buffers (e.g., FPCA PathCov = 0.851 vs. 0.90 nominal, compared with 0.883 under pooled calibration). Accordingly, we view Mondrian conformal primarily as a transparent conditional-calibration diagnostic and a pragmatic enhancement when regime-wise interpretability is a priority, rather than a universal fix for dependence-induced undercoverage.

Figures 1–3 illustrate representative forecast curves and their rolling Mondrian bands. DNS

typically requires visibly wider bands, consistent with its larger point-forecast errors and more heterogeneous residual behavior across maturities. Figure 4 shows that pointwise coverage under Mondrian remains generally conservative across maturities, while Figure 5 confirms that DNS bands are substantially wider, especially in the mid-maturity region. Figure 6 illustrates time-variation in rolling conformal thresholds, with spikes during turbulent periods.

The band width at maturity τ_g satisfies $U_t(\tau_g) - L_t(\tau_g) = 2\, q_t\, s_g(t)$ under studentized rolling conformal prediction. Hence, localized heteroskedasticity or maturity-specific misspecification can inflate widths over a subset of maturities even when differences in *integrated* loss (mISE) appear moderate. In our application, DNS residual curves are more heterogeneous across maturities, producing larger scale estimates $s_g(t)$ (and therefore wider bands) particularly over intermediate maturities, whereas the conformal quantile q_t is not uniformly larger than for FPCA/RawPC.

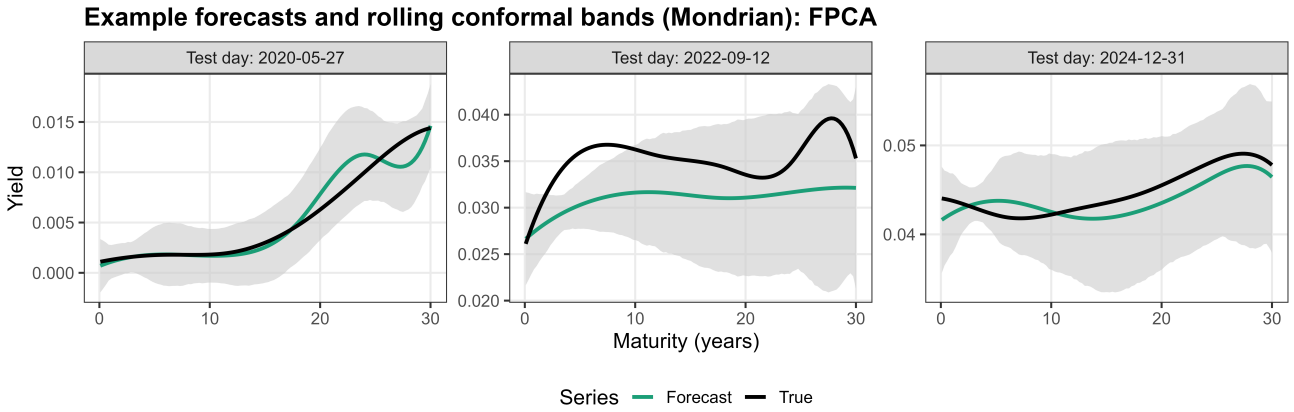

Figure 1. Example forecasts and rolling Mondrian conformal bands: FPCA–VAR (TEST).

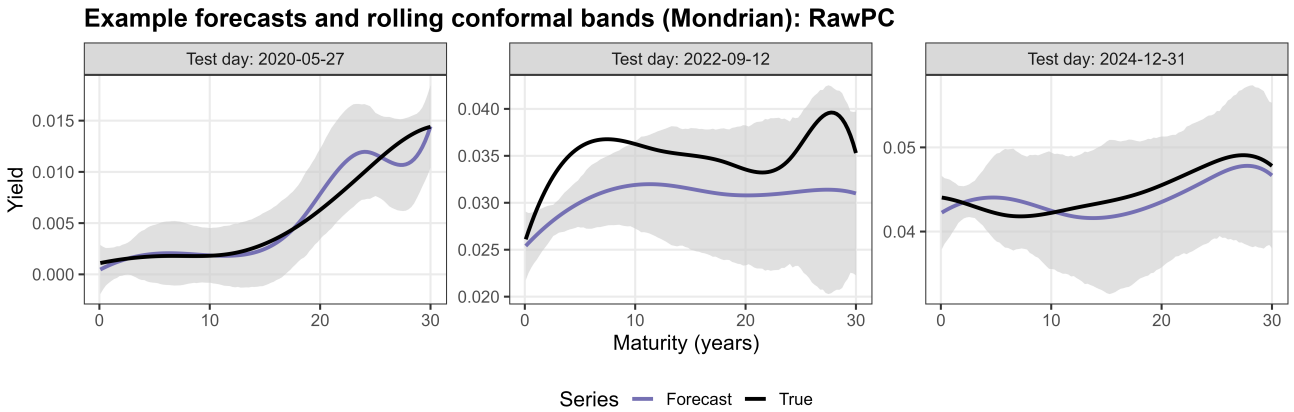

Figure 2. Example forecasts and rolling Mondrian conformal bands: RawPC–VAR (TEST).

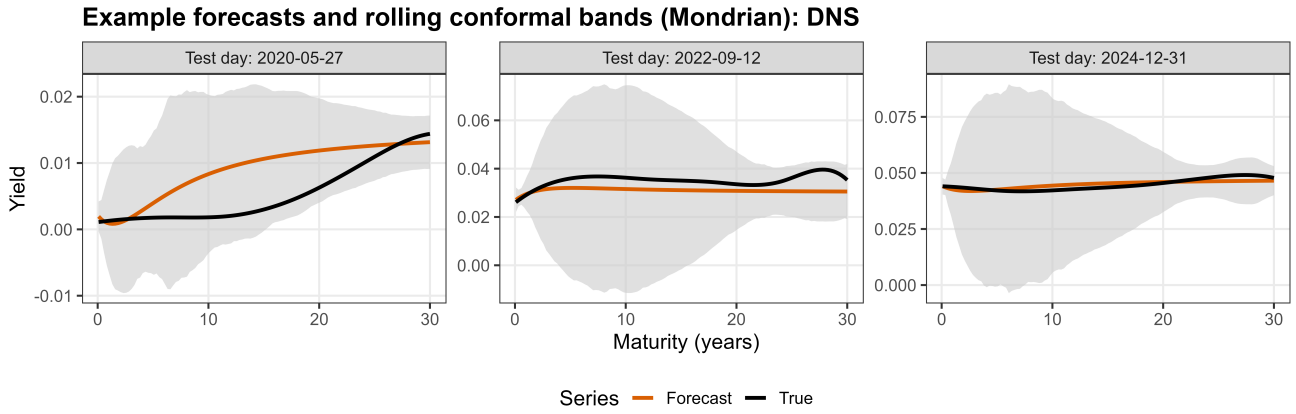

Figure 3. Example forecasts and rolling Mondrian conformal bands: DNS (TEST).

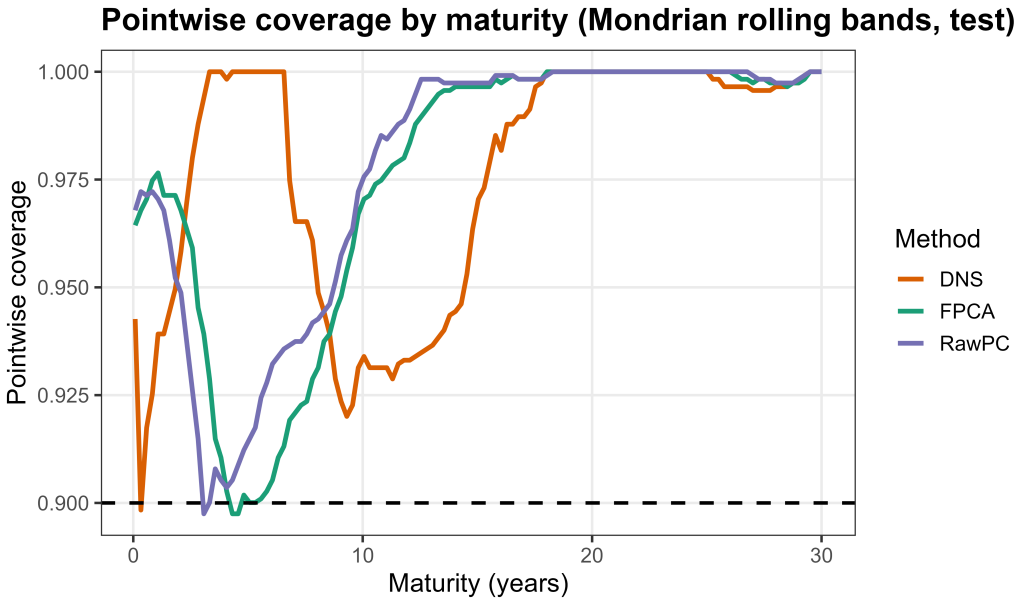

Figure 4. Pointwise coverage by maturity under rolling Mondrian conformal bands (TEST).

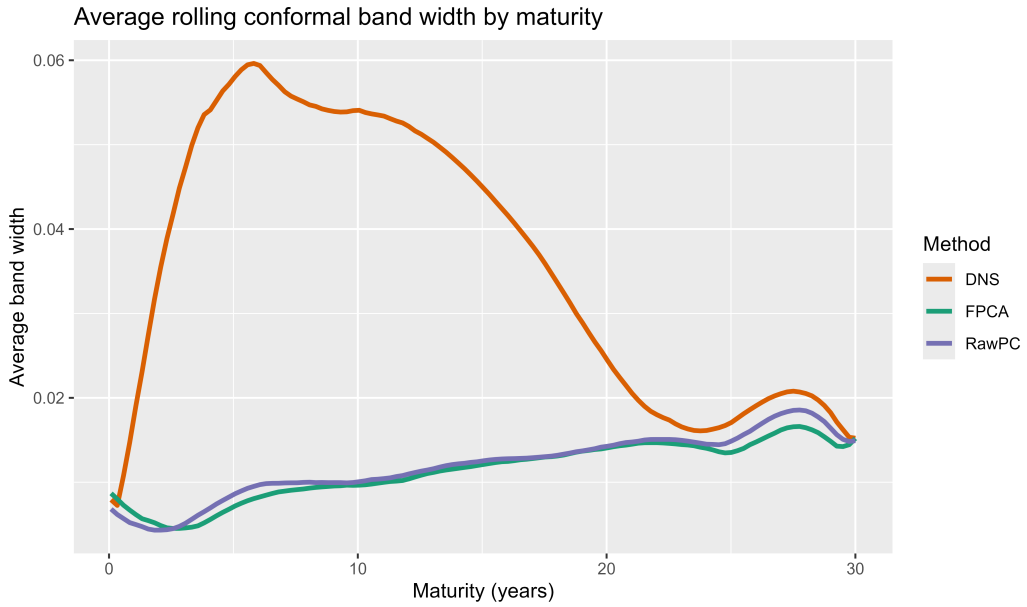

Figure 5. Average conformal band width by maturity under rolling Mondrian bands (TEST).

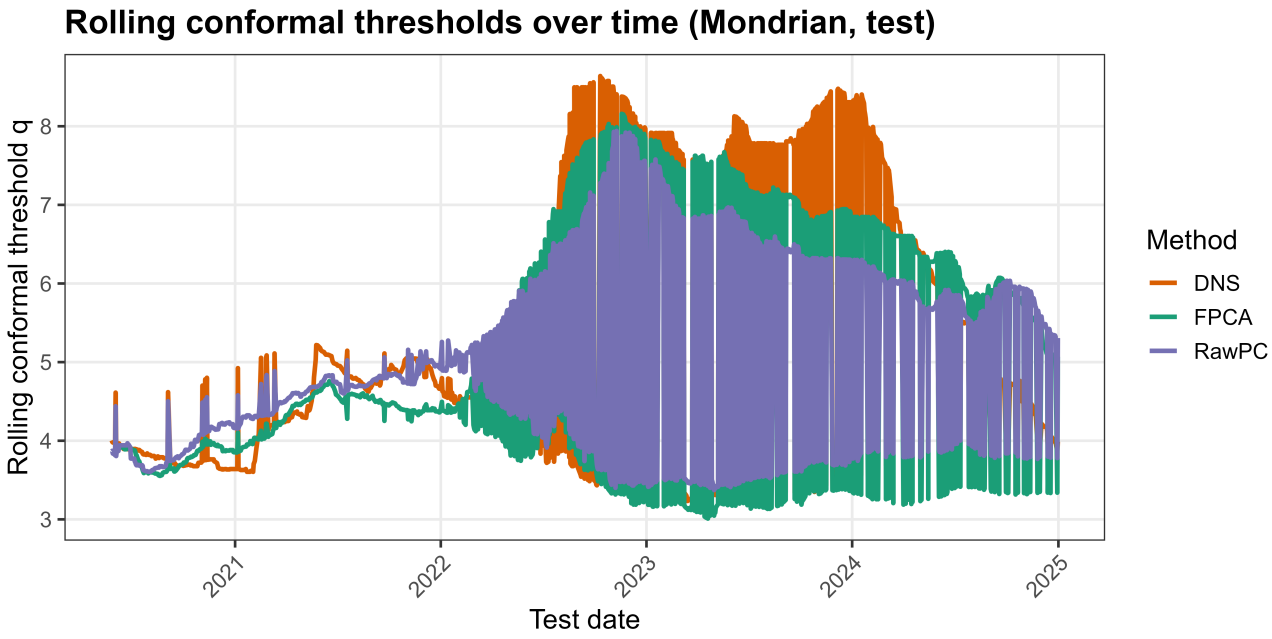

Figure 6. Rolling conformal thresholds over time under the Mondrian scheme (TEST).

5.3. Pairwise significance: DM tests with multiple-testing correction

Table 4 reports Diebold–Mariano tests on daily ISE losses with HAC standard errors (Newey–West lag = $\lfloor T_{\text{te}}^{1/3} \rfloor = 10$), together with Holm and BH/FDR adjustments. All pairwise comparisons are statistically decisive: Both FPCA and RawPC dominate DNS with large test statistics and very small adjusted p-values. Moreover, the FPCA advantage over RawPC, while numerically modest, remains statistically significant after multiple-testing correction.

Table 4. Diebold–Mariano tests on daily ISE losses (TEST), with HAC Newey–West standard errors and multiple-testing correction.

Comparison	MeanDiff	DM	p	Holm	BH	lag
FPCA – DNS	−0.000830845	−13.950	< 0.01	< 0.01	< 0.01	10
RawPC – DNS	−0.000826045	−13.773	< 0.01	< 0.01	< 0.01	10
FPCA – RawPC	−0.0000048002	−3.481	< 0.01	< 0.01	< 0.01	10

5.4. Dependence-robust uncertainty on performance gaps: block bootstrap CIs

To quantify dependence-robust uncertainty in average loss gaps, Table 5 reports moving-block bootstrap confidence intervals for mean ISE differences at canonical block lengths. The qualitative ranking is robust: The FPCA and RawPC improvements over DNS remain strongly negative across block lengths, and the smaller FPCA–RawPC gap remains below zero. Figure 7 shows that larger blocks (preserving more dependence) widen intervals as expected, while leaving sign and ranking conclusions unchanged.

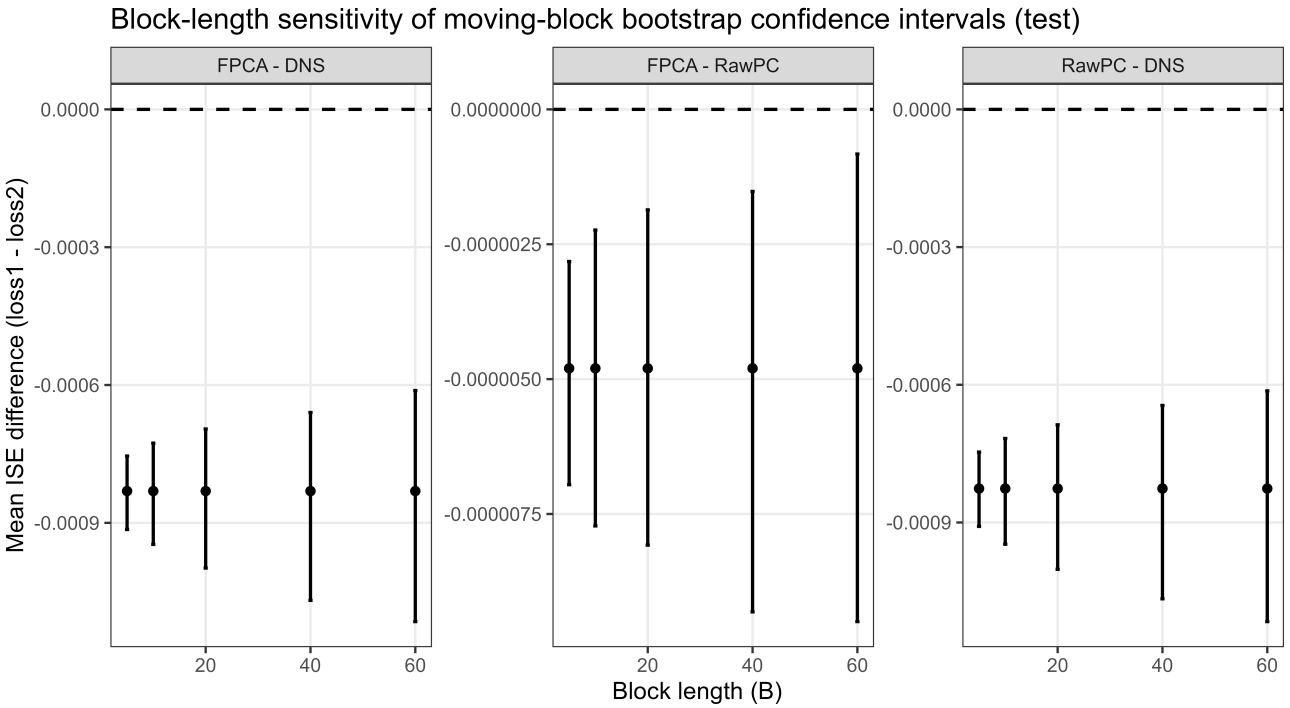

Figure 7. Block-length sensitivity of moving-block bootstrap confidence intervals for mean ISE differences (TEST).

Table 5. Moving-block bootstrap 95% confidence intervals for mean ISE differences on the TEST block.

Contrast	**MeanDiff**	**95% CI** ($R = 2000$)
Block length $B = 5$		
FPCA – DNS	−0.000830845	[−0.000914417, −0.000754645]
RawPC – DNS	−0.000826045	[−0.000908496, −0.000746670]
FPCA – RawPC	−0.0000048002	[−0.00000695721, −0.00000282005]
Block length $B = 10$		
FPCA – DNS	−0.000830845	[−0.000946886, −0.000726532]
RawPC – DNS	−0.000826045	[−0.000947376, −0.000716989]
FPCA – RawPC	−0.0000048002	[−0.00000772149, −0.00000223640]

5.5. Sensitivity and ablations

We next examine how calibration choices affect coverage–width trade-offs. Table 6 varies the rolling conformal window length W (pooled FPCA). Shorter windows react faster but yield lower curve-wise coverage (notably under-coverage for $W = 252$), whereas $W = 750$ provides the best overall balance in this grid. Table 7 demonstrates that both rolling calibration and studentization are essential. Fixed-window approaches fail to adapt and severely undercover in the curve-wise sense. Rolling improves coverage markedly, and adding MAD-based studentization further increases pathwise coverage.

Table 6. Rolling-window sensitivity for studentized rolling conformal bands (FPCA, pooled, TEST). The selected baseline window is shown in bold.

Window W	**PathCov**	**PtCov**	**AvgBW**	$\bar{q}$
252	0.801	0.982	0.0137575	5.4393
500	0.851	0.978	0.0134985	5.6316
750	**0.883**	**0.982**	**0.0126032**	**5.5650**

Table 7. Conformal ablation study for FPCA (TEST).

Variant	**PathCov**	**PtCov**	**AvgBW**
Fixed calibration			
Unstudentized	0.571	0.906	0.00703963
Studentized	0.554	0.913	0.00733254
Rolling calibration			
Unstudentized	0.816	0.973	0.00954734
Studentized	**0.883**	**0.982**	**0.01260320**

5.6. Model confidence set

Table 8 reports the MCS elimination path under ISE loss using moving-block bootstrap resampling. DNS is eliminated first, followed by RawPC, leaving FPCA as the sole model in the final confidence set (Table 9).

Table 8. MCS elimination steps via moving-block bootstrap under ISE loss (TEST).

Step	Action	$T_{\max}$	p	Summary
1	Drop DNS	14.702	< 0.001	DNS is eliminated first under $T_{\max}$, indicating the weakest performance under ISE loss.
2	Drop RawPC	3.530	< 0.001	RawPC–VAR is eliminated next, leaving FPCA–VAR as the sole model in the final MCS.

Table 9. Final MCS at level $\alpha_{\mathrm{MCS}} = 0.10$ under ISE loss (TEST).

Item	Value	Notes
Surviving set	**{FPCA–VAR}**	FPCA–VAR is the only model retained after the elimination sequence.
Eliminated models	{DNS, RawPC–VAR}	These models are removed in Steps 1–2; see Table 8.

5.7. Conditional performance by curve-shock regime

We conclude with regime-conditional diagnostics based on the ex-ante curve-shock split. Table 10 shows that pooled calibration yields materially different curve-wise coverage across regimes for FPCA and RawPC, while Mondrian calibration (Table 11) reduces this disparity by calibrating within regime.

Table 10. Conditional performance and coverage by curve-shock regime (TEST): pooled rolling conformal.

Model	Regime	mISE	PathCov	AvgBW
FPCA–VAR	LOW	1.057×10^{-4}	0.918	0.01184
FPCA–VAR	HIGH	1.912×10^{-4}	0.787	0.01467
RawPC–VAR	LOW	1.089×10^{-4}	0.923	0.01255
RawPC–VAR	HIGH	2.004×10^{-4}	0.765	0.01503
DNS	LOW	1.029×10^{-3}	0.800	0.03455
DNS	HIGH	7.703×10^{-4}	0.790	0.04845

Table 11. Conditional performance and coverage by curve-shock regime (TEST): Mondrian rolling conformal.

Model	Regime	mISE	PathCov	AvgBW
FPCA–VAR	LOW	1.057×10^{-4}	0.841	0.01019
FPCA–VAR	HIGH	1.912×10^{-4}	0.877	0.01451
RawPC–VAR	LOW	1.089×10^{-4}	0.845	0.01020
RawPC–VAR	HIGH	2.004×10^{-4}	0.890	0.01666
DNS	LOW	1.029×10^{-3}	0.746	0.03633
DNS	HIGH	7.703×10^{-4}	0.848	0.03499

5.8. Temporal diagnostics and nonstationarity-aware interpretation

Finally, Figure 8 provides a compact temporal diagnostic by visualizing monthly average absolute forecast errors for FPCA–VAR. The resulting clustering pattern is consistent with time-varying volatility and gradual distributional drift in daily yield-curve dynamics. Together with the rolling threshold dynamics in Figure 6, these diagnostics support a nonstationarity-aware interpretation of conformal validity: Rather than assuming exchangeability, we treat nominal coverage as a practical target and empirically audit calibration over time (Section 4.6).

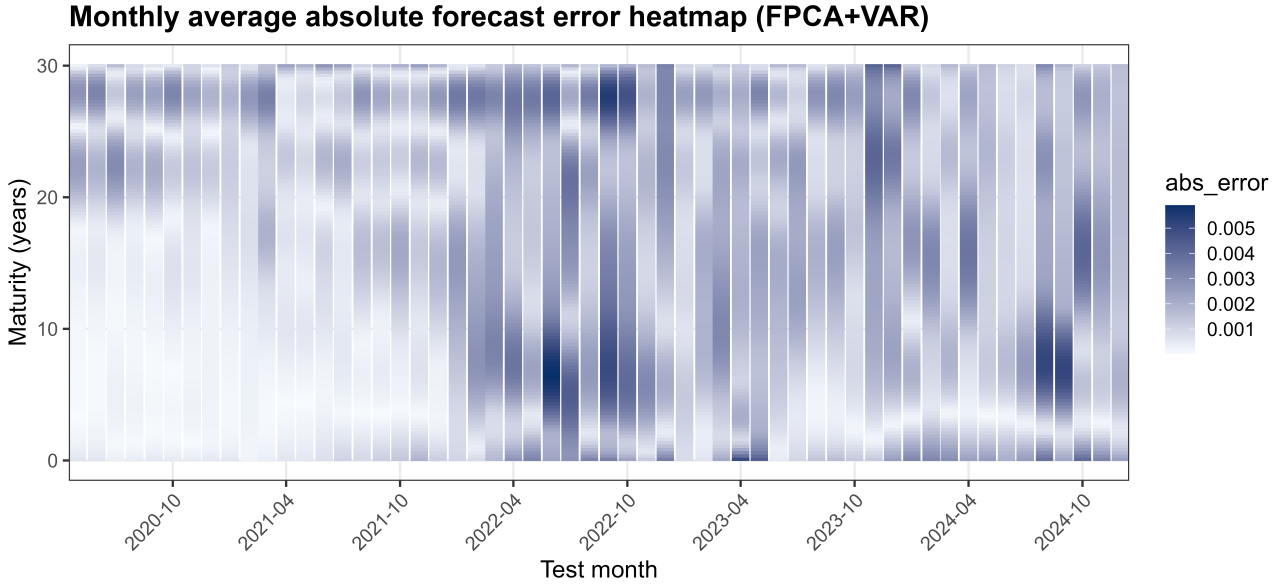

Figure 8. Monthly average absolute forecast error heatmap for FPCA–VAR (TEST).

Empirically, FPCA–VAR consistently achieves the best one-step-ahead daily accuracy on a common maturity grid, and dependence-aware inference (HAC DM tests, moving-block bootstrap CIs, and MCS) confirms that this ranking is robust on TEST. Rolling studentized conformal bands provide a transparent, model-agnostic, distribution-free uncertainty layer for the entire curve, with conservative pointwise coverage arising naturally from simultaneous calibration. Mondrian (shock-conditional) calibration reveals that pooled averages can mask regime-dependent miscalibration and improves interpretability by explicitly auditing calibration across LOW/HIGH shock regimes. At the same time, curve-wise coverage remains sensitive to temporal dependence and finite calibration buffers, which motivates a careful discussion of validity and interpretation under nonstationarity.

6. Discussion

Our out-of-sample evidence delivers a consistent ranking for one-step-ahead daily yield-curve forecasting on a common maturity grid. FPCA–VAR achieves the best integrated accuracy, RawPC–VAR is a close second, and DNS is substantially worse (Table 2). This ordering is stable under dependence-robust evaluation: HAC-corrected Diebold–Mariano tests reject equal predictive accuracy in all pairwise comparisons (Table 4), moving-block bootstrap confidence intervals preserve the same sign and ordering across block lengths, and the MCS retains FPCA–VAR as the unique surviving model

under ISE loss at level $\alpha_{\mathrm{MCS}} = 0.10$.

On uncertainty quantification, rolling studentized conformal bands provide a practical curve-wide calibration layer with directly auditable out-of-sample coverage. Under pooled calibration, curve-wise coverage for FPCA–VAR and RawPC–VAR is close to the nominal level, while pointwise coverage is conservative by construction due to the simultaneous (subtype) score (Table 2). The ablation analysis confirms that both design elements are necessary: Rolling calibration improves stability under gradual drift, and MAD-based studentization mitigates maturity-dependent and time-varying dispersion; fixed-window variants exhibit materially poorer curve-wise calibration (Table 7). The time series of rolling thresholds provides an interpretable diagnostic, expanding during turbulent periods and contracting in calmer regimes.

A central operational takeaway concerns bandwidth drivers. Because width satisfies $U_t(\tau_g) - L_t(\tau_g) = 2\, q_t\, s_g(t)$ under the studentized construction, method-level differences in average width can be driven predominantly by maturity-local scale variation $s_g(t)$ rather than by systematically larger conformal thresholds q_t. In our application, DNS yields substantially wider bands mainly due to more heterogeneous residual dispersion across maturities on the grid, consistent with the pronounced maturity-wise width profile.

Beyond pooled validity, the ex-ante Mondrian design is most informative as a conditional-calibration audit. Pooled calibration can mask regime heterogeneity: FPCA–VAR and RawPC–VAR are more conservative in LOW-shock periods and exhibit weaker curve-wise coverage in HIGH-shock periods. Mondrian calibration reduces this regime imbalance by estimating thresholds within regime without altering the point-forecast ranking. Overall curve-wise coverage may remain below nominal due to temporal dependence and finite within-regime buffers; hence, Mondrian should be viewed as an interpretable diagnostic and pragmatic refinement rather than a guarantee restoration mechanism.

First, classical exchangeability does not hold exactly for daily yields; we therefore treat nominal conformal validity as an operational benchmark and complement coverage diagnostics with dependence-robust inference (HAC and block bootstrap). Second, the HIGH/LOW partition is deliberately simple to preserve stable buffers; richer partitions may sharpen conditional insights but require explicit sample-size control. Third, we focus on one-step-ahead daily forecasts on a fixed grid. Natural extensions include:

(1) **Adaptive conditional calibration:** Multi-bin or data-adaptive Mondrian schemes driven by multiple shock features (e.g., level/slope/curvature changes) with explicit buffer constraints;

(2) **Alternative functional scores:** Nonconformity scores with differential weighting across maturities or target economically relevant curve regions while retaining curve-wide interpretability;

(3) **Richer dynamics:** Incorporating macro covariates or regime-switching structure to better track time variation in residual scale and improve band efficiency.

7. Conclusions

This paper presents a unified, leak-safe framework for one-step-ahead daily yield-curve forecasting together with curve-wide uncertainty quantification on a common maturity grid. After mapping observed maturities to a dense grid via ridge-regularized B-spline smoothing, we compare FPCA–VAR, RawPC–VAR, and a DNS, and equip each with rolling studentized conformal prediction bands.

Empirically, FPCA–VAR delivers the strongest out-of-sample performance. It attains the lowest mISE and produces the tightest conformal bands, while RawPC–VAR remains competitive but statistically distinguishable. DNS is substantially less accurate and yields markedly wider bands, primarily reflecting larger maturity-local residual dispersion on the common grid rather than uniformly larger conformal thresholds. Dependence-aware comparisons—HAC DM tests, moving-block bootstrap confidence intervals, and the MCS procedure—consistently support FPCA–VAR as the leading method under ISE loss.

Rolling studentized conformal bands provide an interpretable, model-agnostic uncertainty layer whose calibration can be audited out-of-sample. Rolling calibration and robust studentization are both essential: Fixed-window alternatives exhibit substantially weaker curve-wise calibration, whereas rolling studentized bands achieve curve-wise coverage close to nominal with conservative pointwise behavior and moderate average width. Mondrian (shock-conditional) calibration further improves interpretability by revealing and partially reducing regime-dependent miscalibration across LOW/HIGH curve-shock conditions. Overall, the results recommend FPCA–VAR paired with rolling studentized conformal bands as a practical daily baseline for yield-curve forecasting with transparent, curve-wide uncertainty assessment.

Appendix

A. Theoretical notes on rolling calibration and block resampling

This appendix briefly records the theoretical rationale behind the two design choices that are most sensitive to temporal dependence: (i) rolling conformal calibration and (ii) block-resampling inference for mean loss gaps. These notes are included for transparency and are not used to claim exact finite-sample guarantees under general time-series dependence.

A.1. Rolling calibration as a local-stability device

In non-exchangeable environments, conformal calibration may deviate from nominal coverage, but recent results provide conditions under which the deviation can remain controlled and small in terms of a measure of dependence or departure from exchangeability; see, e.g., [31] and [32]. Our rolling-window implementation can be viewed as a pragmatic local-stability device: By calibrating on the most recent residual curves, the empirical distribution of nonconformity scores is encouraged to be approximately stable over moderate windows, consistent with locally stationary interpretations of daily yields [34].

Separately, adaptive/online conformal perspectives motivate sliding calibration under distribution shift as a way to track evolving uncertainty in real time (without retraining the forecaster itself), which aligns with the operational role of our rolling buffer [29, 30]. In the main text we therefore adopt a benchmark–audit interpretation: Nominal coverage is a target, and any miscalibration is made explicit via test-block curve-wise and pointwise coverage diagnostics.

A.2. Why block resampling for dependent loss series

For comparative evaluation, daily loss differentials are serially dependent. Consequently, i.i.d. bootstrap resampling is inappropriate because it destroys temporal dependence and typically

understates uncertainty. Block bootstrap methods preserve local dependence by resampling contiguous segments; classical results establish consistency for means (and mean loss gaps) under weak dependence / mixing-type conditions for stationary time series [7, 8]. Our moving-block implementation is therefore a standard dependence-preserving choice for uncertainty quantification of average performance gaps. As an additional audit, we report block-length sensitivity to demonstrate that qualitative conclusions are not an artifact of a single resampling design.

Author contributions

Mervenur Sözen: Data curation, formal analysis, software, visualization, writing–original draft; Fikriye Kabakcı: Conceptualization, methodology, supervision, validation, writing–review and editing; Çağlar Sözen: Conceptualization, methodology, formal analysis, software, writing–review and editing. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

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

Acknowledgments

This study has been supported by the Recep Tayyip Erdogan University Development Foundation (Grant number: 02026001015050).

Conflict of interest

The authors declare no competing interests.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. 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
2. G. Shafer, V. Vovk, A tutorial on conformal prediction, *J. Mach. Learn. Res.*, **9** (2008), 371–421.
3. V. Vovk, A. Gammerman, G. Shafer, *Algorithmic learning in a random world*, Springer, 2005.
4. H. Boström, U. Johansson, T. Löfström, *Mondrian conformal predictive distributions*, In: Proceedings of the 11th Symposium on Conformal and Probabilistic Prediction and Applications (COPA), **152** (2021), 24–38.

5. 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
6. P. R. Hansen, A. Lunde, J. M. Nason, The model confidence set, *Econometrica*, **79** (2011), 453–495. https://doi.org/10.3982/ECTA5771
7. H. R. Künsch, The jackknife and the bootstrap for general stationary observations, *Ann. Stat.*, **17** (1989), 1217–1241. https://doi.org/10.1214/aos/1176347265
8. S. N. Lahiri, *Resampling methods for dependent data*, Springer, 2003.
9. 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
10. Ç. Sözen, Uniform one-sided conformal bands for forward realized volatility curves, *AIMS Math.*, **10** (2025). https://doi.org/10.3934/math.20251201
11. Ç. Sözen, F. Kabakci, Forecasting future realized variance paths with depth-weighted ridge and conformal diagnostics, *AIMS Math.*, **10** (2025), 30246–30270. https://doi.org/10.3934/math.20251329
12. R. Litterman, J. Scheinkman, Common factors affecting bond returns, *J. Fixed Income*, **1** (1991), 54–61. https://doi.org/10.3905/jfi.1991.692347
13. C. R. Nelson, A. F. Siegel, Parsimonious modelling of yield curves, *J. Bus.*, **60** (1987), 473–489. https://doi.org/10.1086/296409
14. L. E. O. Svensson, *Estimating and interpreting forward interest rates: Sweden 1992–1994*, Sveriges Riksbank Working Paper, 1994.
15. F. X. Diebold, C. Li, Forecasting the term structure of government bond yields, *J. Econometrics*, **130** (2006), 337–364. https://doi.org/10.1016/j.jeconom.2005.03.005
16. G. R. Duffee, Term premia and interest rate forecasts in affine models, *J. Finance*, **57** (2002), 405–443. https://doi.org/10.1111/1540-6261.00426
17. J. H. E. Christensen, F. X. Diebold, G. D. Rudebusch, The affine arbitrage-free class of Nelson–Siegel term structure models, *J. Econometrics*, **164** (2022), 4–20. https://doi.org/10.1016/j.jeconom.2011.02.011
18. J. O. Ramsay, B. W. Silverman, *Functional data analysis*, Springer, 2 Eds., 2005.
19. L. Horváth, P. Kokoszka, *Inference for functional data with applications*, Springer, 2012.
20. R. J. Hyndman, H. L. Shang, Forecasting functional time series, *J. Korean Stat. Soc.*, **38** (2009), 199–221. https://doi.org/10.1016/j.jkss.2009.06.002
21. R. J. Hyndman, M. S. Ullah, Robust forecasting of mortality and fertility rates: A functional data approach, *Comput. Stat. Data Anal.*, **51** (2007), 4942–4956. https://doi.org/10.1016/j.csda.2006.07.028
22. S. Hays, H. Shen, J. Z. Huang, Functional dynamic factor models with application to yield curve forecasting, *Ann. Appl. Stat.*, **6** (2012), 870–894. https://doi.org/10.1214/12-AOAS551

23. L. Horváth, P. Kokoszka, J. VanderDoes, S. Wang, Inference in functional factor models with applications to yield curves, *J. Time Ser. Anal.*, **43** (2022), 872–894. https://doi.org/10.1111/jtsa.12642
24. H. L. Shang, F. Kearney, Dynamic functional time-series forecasts of foreign exchange implied volatility surfaces, *Int. J. Forecasting*, **38** (2022), 1025–1049. https://doi.org/10.1016/j.ijforecast.2021.07.011
25. T. H. Khoo, I. M. Dabo, D. Pathmanathan, S. Dabo-Niang, Generalized functional dynamic principal component analysis, *arXiv preprint*, 2024. https://doi.org/10.48550/arXiv.2407.16024
26. E. Diquigiovanni, S. Fontana, S. Vantini, Conformal prediction bands for multivariate functional data, *J. Multivariate Anal.*, **189** (2022), 104879. https://doi.org/10.1016/j.jmva.2021.104879
27. C. Xu, Y. Xie, *Conformal prediction intervals for dynamic time-series*, In: Proceedings of the 38th International Conference on Machine Learning (ICML), **139** (2021), 11559–11569.
28. C. Xu, Y. Xie, Conformal prediction for time series, *IEEE T. Pattern Anal.*, **45** (2023), 11575–11587. https://doi.org/10.1109/TPAMI.2023.3272339
29. I. Gibbs, E. J. Candès, *Adaptive conformal inference under distribution shift*, In: Advances in Neural Information Processing Systems (NeurIPS), 2021. https://doi.org/10.48550/arXiv.2106.00170
30. M. Zaffran, V. Féron, Y. Goude, J. Josse, A. Dieuleveut, *Adaptive conformal predictions for time series*, In: Proceedings of the 39th International Conference on Machine Learning (ICML), **162** (2022), 25834–25866.
31. R. I. Oliveira, P. Orenstein, T. Ramos, J. V. Romano, Split conformal prediction and non-exchangeable data, *J. Mach. Learn. Res.*, **25** (2024), 1–38. https://doi.org/10.48550/arXiv.2203.15885
32. R. F. Barber, A. Pananjady, Predictive inference for time series: why is split conformal effective despite temporal dependence? *arXiv preprint*, 2026. https://doi.org/10.48550/arXiv.2510.02471
33. Federal Reserve Bank of St. Louis, *FRED: Federal Reserve Economic Data* (daily U.S. Treasury constant-maturity series), accessed 2025.
34. R. Dahlhaus, Fitting time series models to nonstationary processes, *Ann. Stat.*, **25** (1997), 1–37. https://doi.org/10.1214/aos/1034276620
35. D. N. Politis, J. P. Romano, The stationary bootstrap, *J. Am. Stat. Assoc.*, **89** (1994), 1303–1313. https://doi.org/10.1080/01621459.1994.10476870

AIMS Press