Proceedings of ICFTBA 2026 Symposium: Digital Government, Finance, Accountability, and Sustainable Economic Development
DOI: 10.54254/2754-1169/2026.AB39126

Forecasting the Monthly-Average U.S. Treasury 10Y–2Y Term Spread: A VARX Model with Macroeconomic Controls

Peiyang Yu

School of Economics and Management, East China Normal University, Shanghai, China
10234700437@stu.ecnu.edu.cn

Abstract. This paper aims to figure out whether the monthly-average 10-year minus 2-year U.S. Treasury term spread can be forecast by its yield-curve and extra macroeconomic information. The research uses monthly data from January 1990 to July 2026 and sets 53 recursive origins during this period. Then the spread is forecast one and three months ahead against a driftless random walk, comparing a second-order vector autoregressive (VAR(2)) model in first differences with a second-order vector autoregressive model with exogenous variables (VARX(2)) that adds inflation, the policy rate, the Chicago Board Options Exchange Volatility Index (VIX) and the economic policy uncertainty (EPU) index. VAR reduces root mean squared error (RMSE) by 7.76% at one month and 3.96% at three months, and the one-month gain survives Holm adjustment ($p = 0.0136$); the VARX-versus-VAR increment is insignificant at both horizons, so the macroeconomic contribution is weak and not robust. Ex-post groupings concentrate the gains where the spread actually narrowed, which describe forecast-error heterogeneity only. Conditional on other yield lags and the macro controls, lagged 2-year changes carry no robust incremental linear information. The results are pseudo-out-of-sample because the data used are revised.

Keywords: term spread, VARX model, pseudo-out-of-sample forecasting, Clark–West test

1. Introduction

The topic is the 10Y–2Y term spread of U.S. Treasury securities, chosen for three reasons. First, the U.S. Treasury market's yield curve is the benchmark risk-free term structure and an anchor for asset pricing around the world, so the behaviour of its slope is of general interest. Second, the spread carries information that yield levels do not: it combines the expected path of short-term interest rates with a term premium, and a negative spread has long been treated as a leading signal of recessions [1, 2]. Third, the predictability of the spread depends on the maturity pair chosen [3]. The 10-year and 2-year points are the most actively traded maturities on the curve. Policy uncertainty enters the information set for a related reason: it affects yield levels, volatility and term premia alike [4]. The curve inverted in July 2022 and stayed inverted until August 2024, which gives the recent sample particular interest. Using monthly data over a long sample, the paper forecasts the spread one and three months ahead against a random-walk (RW) benchmark on identical recursive origins, and asks whether macroeconomic information adds a robust increment beyond yield-curve dynamics.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.AB39126

On method, the paper compares two sets of nested linear specifications, macro-augmented vector autoregression with exogenous variables (VARX) against vector autoregression (VAR), and VAR against a driftless random walk. Nested linear specifications suit the research topic well because they can serve as a forecast comparison that separate a system's own dynamics from the increment due to additional information. Also, evaluation is subject to the suitability of methods. Accuracy comparisons between nested models call for the Clark–West (CW) adjusted statistic with Holm adjustment across these primary comparisons [5], but given that heteroskedasticity- and autocorrelation-consistent (HAC) tests perform poorly in finite samples [6] and that test statistics for nested models have non-standard asymptotic distributions [7], the statistics should be read with caution here.

Previous studies provide the data and method reference points for this paper, but their conclusions conflict, and the disagreement points to questions that remain open. Huang et al. [8] and Wan et al. [9] both forecast U.S. Treasury (bond) returns and reach opposite conclusions: the former find predictive power from real-time macro data that is not spanned by the yield curve, while the latter show that evidence built on revised data disappears once real-time information is used, so the disagreement turns on the data vintage rather than the market. Fan et al. [10] and Bianchi et al. [11] also use U.S. Treasury data, the former evaluating excess returns by maturity and the latter finding that unspanned factors vary with maturity. On the uncertainty and maturity-pair dimensions examined here, Leippold and Matthys [4] characterise the response of the U.S. yield curve to policy uncertainty, Kang and Ryu [12] find that individual uncertainty measures have limited predictive power, and Choi et al. [3] show that the choice of maturity pair affects conclusions; Salachas et al. [13] and Tvedt [14] caution that term-structure conclusions are period- and country-sensitive and that predictive power rests on the whole curve. Two questions therefore remain insufficiently answered. First, most studies forecast bond excess returns or yield levels, so comparisons that take the term spread as the target and separate the two contributions are rare. Second, the coexistence of positive and negative evidence shows that conclusions are sensitive to specification and evaluation. The paper accordingly poses three research questions.

RQ1. Do yield dynamics improve one- and three-month-ahead spread forecasts relative to Random Walk, and does VARX improve accuracy further relative to VAR?

RQ2. How do pseudo-out-of-sample forecast errors differ across horizons and across ex-post widening, narrowing, inversion and non-inversion observations?

RQ3. Do lagged 2-year yield changes contain incremental linear predictive information for 10-year yield changes after conditioning on other yield lags and macroeconomic controls?

2. Data and sample construction

The monthly Federal Reserve Economic Data (FRED) panel covers January 1990–July 2026 (439 months) [15]; Table 1 lists the definitions and monthly construction. The initial training sample is January 1990–December 2021 (384 months). Forecasts are updated over 53 common origins, December 2021–April 2026, for both horizons and all models. Targets are January 2022–May 2026 for h = 1 and March 2022–July 2026 for h = 3. The final two available one-month origins are excluded to keep the origin set identical across horizons. The incomplete August 2026 observation is excluded.

The seasonally adjusted Consumer Price Index (CPI) is published by the U.S. Bureau of Labor Statistics. FEDFUNDS is a monthly series. The monthly Chicago Board Options Exchange Volatility Index (VIX) and economic policy uncertainty (EPU) index inputs are arithmetic means of available daily observations.

DOI: 10.54254/2754-1169/2026.AB39126

Year-on-year inflation is calculated as $\pi_t = 100(\frac{\mathrm{CPI}_t}{\mathrm{CPI}_{t-12}} - 1)$, where CPI is the seasonally adjusted index. The primary model input is $\Delta\pi$, its month-to-month change, in percentage points (pp). The spread equals the monthly-average 10-year yield minus the monthly-average 2-year yield.

October 2025 CPI is missing following the federal shutdown. Year-on-year inflation is filled using a one-step first-order autoregressive (AR(1)) prediction based only on earlier observations, before its first difference is constructed. All calendar months are retained, and RW, VAR and VARX use identical forecast origins and targets. Carry-forward filling is examined as a CPI sensitivity check.

Table 1. Data definitions and monthly construction

Variable / series	Provider / frequency	Monthly construction	Model input
Treasury yields DGS2, DGS5, DGS10	Federal Reserve Board Business Daily	Arithmetic mean of observed daily yields	First difference; pp
Inflation CPIAUCSL	U.S. Bureau of Labor Statistics Monthly; Seasonally Adjusted	$\pi_t = 100(\frac{\mathrm{CPI}_t}{\mathrm{CPI}_{t-12}} - 1)$	First difference of year-on-year inflation; pp
Effective policy rate FEDFUNDS	Federal Reserve Board Monthly	Published monthly effective federal funds rate	First difference; pp
Volatility VIXCLS	Chicago Board Options Exchange Business Daily	Arithmetic mean of observed daily index values	First log difference
Policy uncertainty USEPUINDXD	Baker, Bloom and Davis Calendar Daily	Arithmetic mean of observed daily index values	First log difference

Figure 1 shows that the three monthly-average yields move broadly together, declining from the 1990s to around 2020 and rising markedly from 2022, while the term spread turns negative several times, most recently staying inverted from July 2022 to August 2024. Monthly-average yields are actually arithmetic means of available daily FRED observations in each month, and the term spread equals monthly-average 10-year minus 2-year yields.

DOI: 10.54254/2754-1169/2026.AB39126

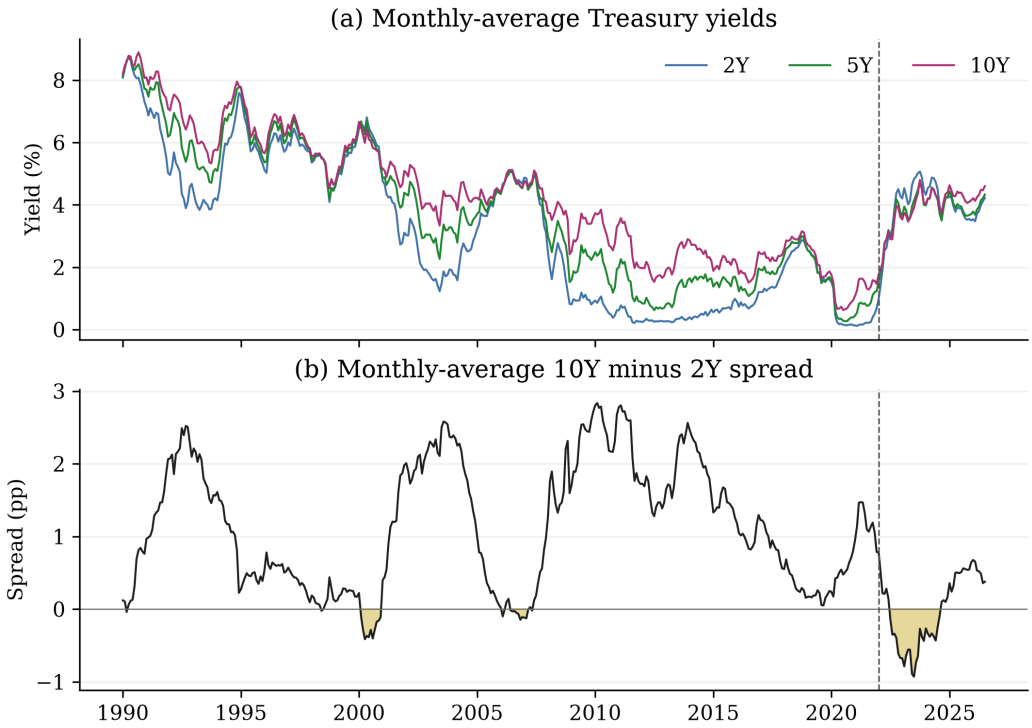


Figure 1. Monthly-average yields and the 10Y–2Y spread

Table 2 reports descriptive statistics, including standard deviations (SD), for January 1990–July 2026. The mean 10Y–2Y spread is 1.001 percentage points, ranging from −0.929 to 2.834. Inflation has 438 observations, while the other series have 439.

Table 2. Descriptive statistics, January 1990–July 2026

Series	N	Mean	SD	Min.	Max.
2Y yield (%)	439	3.251	2.252	0.115	8.724
5Y yield (%)	439	3.764	2.091	0.267	8.768
10Y yield (%)	439	4.252	1.923	0.624	8.893
10Y–2Y (pp)	439	1.001	0.909	−0.929	2.834
Inflation (%)	438	2.710	1.588	−1.959	8.979
Federal funds (%)	439	2.891	2.326	0.050	8.290
VIX (index)	439	19.453	7.367	10.125	62.669
EPU (index)	439	122.127	79.085	38.209	560.876

3. Empirical methodology

3.1. Model, estimation and forecast target

Let $\mathbf{y}_t = (y_{2,t},\ y_{5,t},\ y_{10,t})'$ and $\mathbf{x}_t = (\Delta\pi_t,\ \Delta f_t,\ \Delta\ln \mathrm{VIX}_t,\ \Delta\ln \mathrm{EPU}_t)'$, where π denotes year-on-year inflation and f is the effective federal funds rate. The working model is estimated

DOI: 10.54254/2754-1169/2026.AB39126

equation by equation by ordinary least squares (OLS), with an intercept, two yield lags and one lag of each control, that is, Equation (1):

$$\Delta y_t \;=\; c \;+\; A_1 \Delta y_{t-1} \;+\; A_2 \Delta y_{t-2} \;+\; B x_{t-1} \;+\; \varepsilon_t \tag{1}$$

The matched VAR sets B $=$ 0 and uses the same response dates and two yield lags. Both initial models have 381 estimation observations. Estimation uses an expanding window and is repeated at every origin. This specification follows the standard multivariate VAR framework.

At each forecast origin t, the yield system is estimated using yield data through t, while the four transformed controls are available only through t − 1. A separate AR(1) with an intercept is fitted by OLS to each control using its expanding history through t − 1, after the stated CPI filling rule. Equation (2) recursively predicts the controls for t, t + 1 and t + 2. These predictions enter the yield-change forecasts for t + 1, t + 2 and t + 3, respectively.

$$\widehat{x}_{k,t+j|t} \;=\; \widehat{a}_k \;+\; \widehat{\varphi}_k \, \widehat{x}_{k,t+j-1|t}, \quad j \;=\; 0,1,2 \tag{2}$$

The recursion starts from $\widehat{x}_{k,t-1|t} \;=\; x_{k,t-1}$. Here k indexes the four transformed controls, and the estimated AR(1) intercept and slope are specific to each forecast origin and held fixed during its recursion. For a one-month yield forecast, only the prediction for month t is required; for a three-month forecast, predictions for t, t + 1 and t + 2 are used in sequence. No actual macroeconomic observations dated t or later enter these forecasts. The exercise is pseudo-out-of-sample because the historical series are revised rather than real-time vintages.

Forecasted yield changes are accumulated separately for each maturity following Equation (3); the forecast spread is then formed from the reconstructed monthly-average yield levels:

$$\hat{y}_{m,t+h|t} \;=\; y_{m,t} \;+\; \textstyle\sum_{j=1}^{h} \Delta \hat{y}_{m,t+j|t}, \quad m \;\in\; \{2,5,10\} \tag{3}$$

$$\hat{s}_{t+h|t} \;=\; \hat{y}_{10,t+h|t} \;-\; \hat{y}_{2,t+h|t}, \quad h \;\in\; \{1,3\} \tag{4}$$

Random Walk without drift forecasts each yield at the point of its latest monthly average, $\hat{s}^{\mathrm{RW}}_{t+h|t} \;=\; s_t$. All models in equation (4) are evaluated on the spread level. At the three-month horizon, the target is the spread in month t + 3.

3.2. Specification and predictive-accuracy tests

The time period of tests is January 1990–December 2021. As unit-root tests, augmented Dickey–Fuller (ADF) tests include a constant or a constant and trend, with 12 augmentation lags; Phillips–Perron (PP) and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) supply complementary evidence, and Johansen tests use a constant restricted to the cointegrating relation. The primary levels lag $K \;=\; 3$ corresponds to the two yield-change lags to be introduced later. Trace rank is the primary rule, while maximum-eigenvalue tests and other lag orders assist by assessing sensitivity.

The working VAR and VARX specifications contain two yield lags. Two lags are retained as a parsimonious benchmark allowing dynamics beyond one month, despite residual serial correlation. The $p \;=\; 7$ systems provide the principal lag-order robustness check, with order 3 as an adjacent-order comparison; the Bayesian information criterion (BIC) selects one lag and the Akaike information criterion (AIC) selects seven in the initial training sample. CPI sensitivity covers

DOI: 10.54254/2754-1169/2026.AB39126

inflation in levels and carry-forward filling before differencing. The training evidence for these choices is reported in Section 4.

For model q and horizon h, the forecast error is $e_{q,t,h} = s_{t+h} - \hat{s}^{(q)}_{t+h|t}$. Accuracy is measured by mean absolute error (MAE), root mean squared error (RMSE) and the out-of-sample coefficient of determination, as defined in equations (5)–(7):

$$\mathrm{MAE}_{q,h} = \frac{1}{N}\sum_{t\in T}|e_{q,t,h}| \tag{5}$$

$$\mathrm{RMSE}_{q,h} = \sqrt{\frac{1}{N}\sum_{t\in T}e^2_{q,t,h}} \tag{6}$$

$$R^2_{\mathrm{OS},q,h} = 1 - \frac{\sum_{t\in T}e^2_{q,t,h}}{\sum_{t\in T}e^2_{\mathrm{RW},t,h}} \tag{7}$$

The primary comparisons are VAR against RW and VARX against VAR. Clark–West adjusted squared-loss differences test the one-sided alternative that the larger nested model adds predictive accuracy. In equation (8), b denotes the benchmark and a the larger model:

$$d^{\mathrm{CW}}_{t,h} = e^2_{b,t,h} - e^2_{a,t,h} + \left(\hat{s}_{b,t+h|t} - \hat{s}_{a,t+h|t}\right)^2 \tag{8}$$

Inference uses HAC standard errors to allow for heteroskedasticity and serial correlation in the loss differences, with Holm adjustment across the four primary comparisons. Supplementary Diebold–Mariano (DM) tests use squared and absolute loss with the Harvey–Leybourne–Newbold correction. The 53 recursively estimated forecasts and overlapping three-month targets make these tests approximate; DM provides supplementary evidence for the nested comparisons.

RQ3 jointly tests the two lagged 2-year yield-change coefficients in the 10-year yield-change equation, conditional on other yield lags and the macroeconomic controls. HAC covariance is the primary estimate; conventional OLS covariance is also reported. The test evaluates conditional incremental linear predictive information. It has no causal interpretation and is distinct from the out-of-sample forecast comparison.

3.3. Ex-post error classifications

Widening means $s_{t+h} - s_t > 0$ and narrowing means $s_{t+h} - s_t < 0$; unchanged spreads are identified using tolerance 10^{-12}, and none occur. Inversion means $s_{t+h} < 0$ and non-inversion means $s_{t+h} \geq 0$. These are two overlapping classifications based on realized target values. RQ2 therefore describes ex-post forecast-error heterogeneity; it does not assess advance identification of market states or formal state-dependent forecasting.

4. Training-sample evidence

4.1. Stationarity and cointegration tests

As reported in Table 3, tests for year-on-year inflation in levels disagree: ADF does not reject a unit root, PP rejects it, and KPSS rejects stationarity. For differenced inflation, ADF and PP reject a unit root while KPSS does not reject stationarity. This evidence supports differenced inflation in the working model; inflation in levels is retained as a sensitivity specification. Differenced inflation

DOI: 10.54254/2754-1169/2026.AB39126

measures the monthly change in inflation rather than its level, so its coefficient should be read as the marginal effect of inflation changes on spread changes. The primary input is chosen on specification evidence rather than forecast performance: Table 11 shows that inflation in levels actually yields lower realised forecast errors in this sample, so the paper does not claim that differencing necessarily improves accuracy.

Table 3. Unit-root and stationarity tests, January 1990–December 2021

Series	Form	N	ADF(c)	ADF(ct)	PP(c)	KPSS
2Y	Level	384	−2.073	−3.539*	−1.902	4.922*
2Y	Δ	383	−4.489*	−4.490*	−12.528*	0.090
5Y	Level	384	−1.764	−3.505*	−1.768	5.603*
5Y	Δ	383	−5.423*	−5.435*	−13.828*	0.055
10Y	Level	384	−1.563	−4.376*	−1.574	5.942*
10Y	Δ	383	−6.160*	−6.165*	−14.625*	0.029
Spread	Level	384	−3.106*	−3.093	−2.246	0.400
Spread	Δ	383	−4.808*	−4.815*	−14.101*	0.096
CPI YoY	Level	384	−2.818	−2.060	−3.263*	1.137*
ΔCPI YoY	Input	383	−7.869*	−8.088*	−12.013*	0.167
ΔFed funds	Input	384	−4.360*	−4.333*	−10.150*	0.131
Δln VIX	Input	383	−7.214*	−7.217*	−20.085*	0.024
Δln EPU	Input	384	−7.115*	−7.126*	−22.561*	0.016

* denotes rejection of the respective null hypothesis at the 5% level. The ADF and Phillips–Perron null is a unit root; the KPSS null is stationarity. c = constant; ct = constant and linear trend.

Table 4 uses yield levels over January 1990 to December 2021 with the constant restricted to the cointegrating relation. At the levels lag order corresponding to the two-lag differenced specification, K = 3, neither the trace nor the maximum-eigenvalue test rejects the null of no cointegration (trace statistic 28.963 against a critical value of 34.91; maximum-eigenvalue statistic 17.555 against 22), so the baseline evidence supports a first-differenced VAR without an error-correction term. The conclusion is sensitive to the levels lag order: at K = 8, corresponding to the seven-lag specification, both tests still select rank zero, but at K = 9 the maximum-eigenvalue test selects rank one (statistic 22.375, just above the critical value of 22). The appropriate statement is therefore that no cointegration is found under the baseline specification, not that no cointegration exists among the three yields.

Table 3's constant-only ADF rejects a unit root in the fixed 10Y–2Y spread, but adding a trend changes the 5% critical value from −2.87 to −3.42 and removes significance. Johansen instead tests system-wide cointegration with a restricted constant and two difference lags, versus 12 ADF lags. These specification differences can produce divergent finite-sample decisions. The rank-zero result in Table 4 therefore does not rule out cointegration.

DOI: 10.54254/2754-1169/2026.AB39126

Table 4. Johansen tests of no cointegration

Levels K	Effective N	Trace	Max. eigenvalue	Rank: trace / max.
3 (primary)	381	28.963	17.555	0 / 0
2	382	31.122	18.960	0 / 0
4	380	30.814	18.177	0 / 0
6	378	25.446	16.084	0 / 0
8	376	28.895	18.894	0 / 0
9	375	34.418	22.375	0 / 1

4.2. Model diagnostics and conditional predictive information

On the initial training sample BIC selects one yield lag and AIC selects seven, so the seven-lag systems carry the principal lag-order robustness check. Table 5 reports residual tests through 12 lags and the largest companion root for the initial training models. Both $p = 2$ systems show residual serial correlation: the Breusch–Godfrey (BG) and portmanteau (PT) p-values are below 0.001. For the $p = 7$ systems, the BG p-values are 0.1663 and 0.1151, and the portmanteau p-values are 0.3735 and 0.1878. All reported roots are below one. The two-lag specification remains a working model with unresolved residual dynamics; the seven-lag systems provide an important robustness check. Residual dynamics are not fully captured at two lags, whereas the seven-lag systems no longer reject the null, indicating that the serial correlation left over by the two-lag specification can be absorbed by a longer lag structure. The seven-lag systems therefore serve to test the sensitivity of the conclusions to lag structure; whether predictive performance also improves is a separate question answered by the out-of-sample comparison in Section 5. The two-lag specification is retained as a parsimonious working model and is not presented as a final specification that fully passes diagnostics.

Table 5. Training-model diagnostics

Model	p	N	Largest root	BG p	Adjusted PT p
VAR	2	381	0.392	<0.001	<0.001
VARX	2	381	0.415	<0.001	<0.001
VAR	7	376	0.878	0.166	0.373
VARX	7	376	0.908	0.115	0.188

Table 6 tests $H_0: \mathbf{A}_1[10Y,2Y] = \mathbf{A}_2[10Y,2Y] = 0$ in the initial-training VARX(2), with N = 381 and differenced inflation. The HAC Wald statistic is 0.2984 (p = 0.8614); the conventional OLS result is similar (p = 0.8584). Under this conditional linear specification, lagged 2-year yield changes do not provide statistically significant incremental predictive information for 10-year yield changes. The result indicates that, once other yield lags and the four macroeconomic controls are accounted for, lagged 2-year yield changes carry no identifiable incremental linear information for 10-year yield changes. Because the test concerns conditional linear predictive information, it should

DOI: 10.54254/2754-1169/2026.AB39126

not be read as evidence that no economic relationship exists between the two maturities; their contemporaneous and long-run relation lies outside this test.

Table 6. Joint Wald test: lagged 2Y changes in the 10Y equation

Covariance	Wald $\chi 2$	df	p-value	2Y lag 1	2Y lag 2
OLS	0.305	2	0.858	0.022	0.088
HAC12	0.298	2	0.861	0.022	0.088

5. Pseudo-out-of-sample forecast evidence

5.1. Overall forecast performance and accuracy tests

Table 7 evaluates monthly-average spread forecasts on 53 common origins. MAE, RMSE and bias are in percentage points; bias is actual minus forecast. R^2_{OS} compares squared loss with driftless RW. Relative to RW, VAR reduces RMSE by 7.76% at one month and 3.96% at three months, from 0.1365 to 0.1259 and from 0.2651 to 0.2546, respectively. The reductions are modest at both horizons but consistent in sign, and the out-of-sample R^2 is positive, indicating that the joint dynamics of yields do contain usable information about the short-horizon spread. The three-month gain is limited, and a statistically significant improvement need not correspond to an economically exploitable gain; the two should be assessed separately.

VARX further reduces RMSE to 0.1250 and 0.2521, improvements of 0.70% and 0.98% over matched VAR. Its MAE is also lower than VAR at both horizons, at 0.0969 and 0.1907. The overall RMSE reductions relative to RW are 8.41% and 4.89%. The further error reduction contributed by the macroeconomic variables is economically small and statistically insignificant. One interpretation is that the state information reflected in inflation, the policy rate, volatility and policy uncertainty is largely spanned by the joint dynamics of the three yields; an enhanced information set improves forecasts only in specific periods [13], so the stability of the spread signal should be judged sample by sample [2]—consistent with the general understanding that macro factors are informative mainly about the term-premium component of the curve. This is an interpretation of the observed result, not a tested mechanism.

Table 7. Spread forecast accuracy on 53 common origins

h	Model	N	MAE	RMSE	R^2_{OS}	Bias
1	RW	53	0.0989	0.1365	0.0000	−0.0056
1	VAR	53	0.0981	0.1259	0.1493	0.0093
1	VARX	53	0.0969	0.1250	0.1611	0.0116
3	RW	53	0.2078	0.2651	0.0000	−0.0159
3	VAR	53	0.1950	0.2546	0.0776	0.0087
3	VARX	53	0.1907	0.2521	0.0955	0.0128

Average loss can conceal differences in signal timing across origins [16]. Figure 2 compares the realised spread with the three forecast paths.

DOI: 10.54254/2754-1169/2026.AB39126

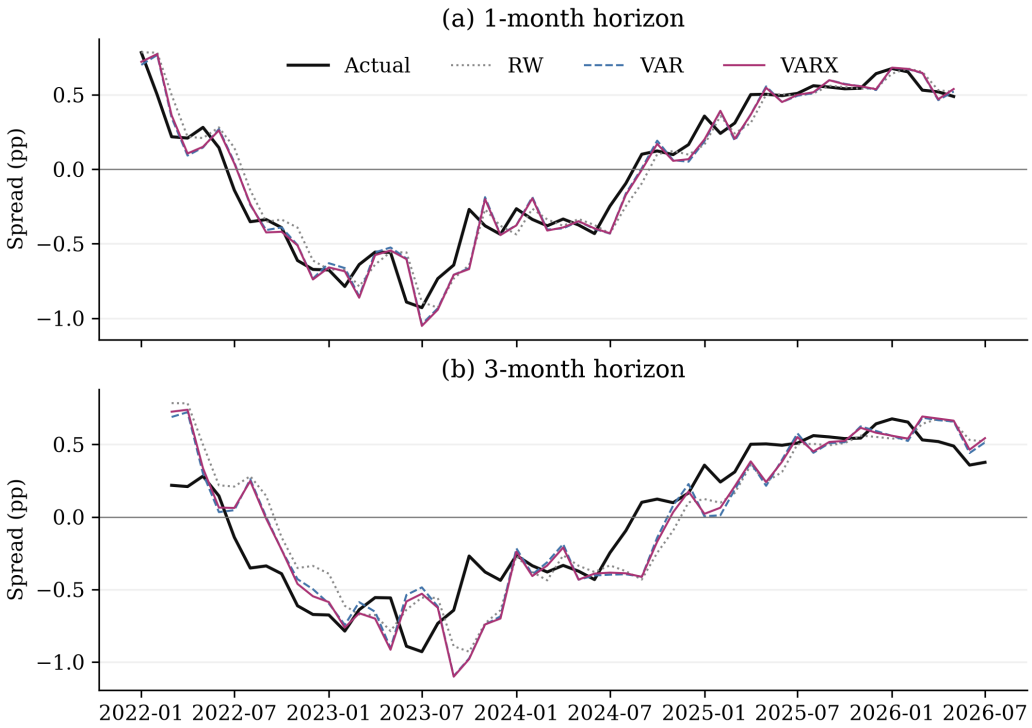


Figure 2. Actual and predicted monthly-average term spreads

Table 8 reports the primary Clark–West tests and supplementary DM tests. After Holm adjustment across four comparisons, the one-month VAR–RW improvement remains significant at 5% (adjusted p = 0.0136). The three-month VAR–RW and both VARX–VAR comparisons are not significant (adjusted p = 0.2333 each). The supplementary DM tests do not confirm the one-month VAR–RW gain at 5%. The primary test therefore supports some short-horizon predictive value in yield dynamics; the additional macroeconomic contribution is weak and lacks robust significance. Only one comparison is statistically supported: the one-month VAR improvement over the random walk. The macroeconomic increment fails the adjusted test at both horizons and is therefore not presented as an improvement in predictive ability. Table 8 concerns system-level forecast comparisons, which differ in object from the RQ3 joint test on specific lagged coefficients; the former does not become invalid because the latter is insignificant.

Table 8. One-sided predictive-accuracy tests

h	Comparison	CW stat.	CW p	Holm p	DM^2 p	DM|e| p
1	VAR vs. RW	2.706	0.003	0.014	0.086	0.447
1	VARX vs. VAR	1.293	0.098	0.233	0.210	0.191
3	VAR vs. RW	1.420	0.078	0.233	0.285	0.225
3	VARX vs. VAR	1.097	0.136	0.233	0.245	0.120

DOI: 10.54254/2754-1169/2026.AB39126

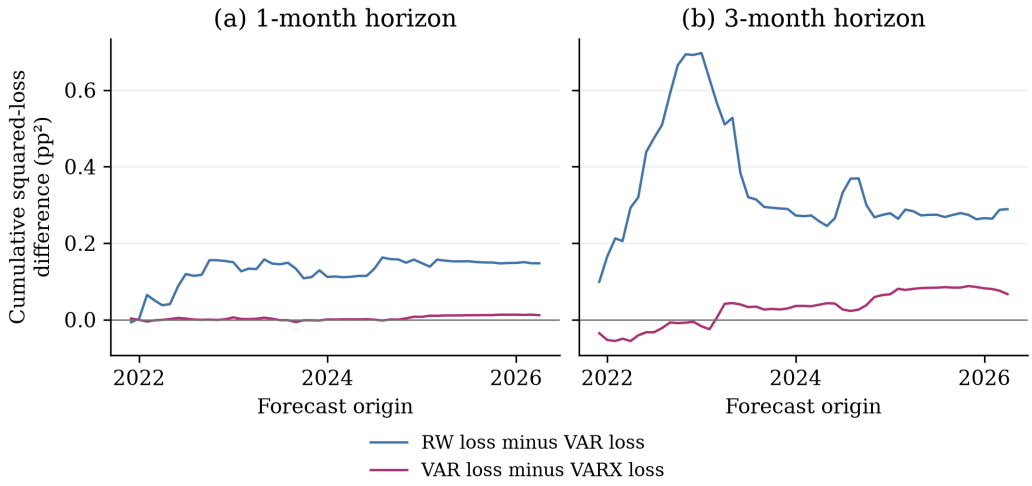


Figure 3. Cumulative raw squared-loss differences by origin

5.2. Ex-post forecast-error heterogeneity

Table 9 reports ex-post subgroup RMSE. Relative to RW, VARX reduces RMSE in narrowing observations by 14.35% at one month and 15.75% at three months. The corresponding VAR reductions are 12.91% and 14.49%. In widening observations, VARX reduces one-month RMSE by 1.62%, but its three-month RMSE is 8.49% higher than RW. One interpretation is that narrowing periods exhibit more pronounced mean reversion, whereas widening periods behave more like trend continuation that the model cannot extrapolate. Because the grouping is defined on realised target-month outcomes, this difference is descriptive ex post and does not support a claim that the model identifies market states in advance.

For every model and horizon, RMSE is higher in inversion observations than in non-inversion observations. These comparisons describe the realized subgroup errors; statistical differences between groups have not been tested. In magnitude, VARX RMSE is 0.2804 in three-month inversion observations against 0.2214 in non-inversion observations, with the gap pointing the same way at both horizons. Within the sample, inversions cluster after rapid increases in the level of rates, a period of larger spread movements, so the higher errors are consistent with the sample background.

Table 9. Ex-post error heterogeneity: RMSE

h	Classification	N	RW	VAR	VARX
1	narrowing	30	0.1346	0.1172	0.1153
1	widening	23	0.1389	0.1364	0.1367
1	inverted	26	0.1566	0.1448	0.1450
1	non-inverted	27	0.1138	0.1045	0.1021
3	narrowing	24	0.3008	0.2572	0.2534
3	widening	29	0.2313	0.2523	0.2510
3	inverted	26	0.2894	0.2867	0.2804
3	non-inverted	27	0.2394	0.2192	0.2214

DOI: 10.54254/2754-1169/2026.AB39126

5.3. Specification sensitivity

Table 10 uses identical forecast origins and differenced inflation across VARX lag specifications. The working order is $p = 2$; $p = 1$ is the training BIC choice, $p = 7$ is the training AIC choice and the main lag robustness check, and $p = 3$ is the adjacent-order comparison. Every reported VAR and VARX has lower RMSE than RW at both horizons. At seven lags, however, VARX RMSE is 0.1301 and 0.2544, compared with 0.1285 and 0.2481 for matched VAR. The two-lag model's numerical macroeconomic increment is not retained at seven lags. The numerical advantage over the random walk holds at every lag order, so the contribution of yield dynamics does not depend on the particular lag choice. The macroeconomic increment, by contrast, is unstable and marginally favourable only at two lags; this contrast echoes the training-sample diagnostics—improved diagnostics do not automatically produce improved forecasts.

Table 10. RMSE sensitivity to yield lag order

p	VAR, h = 1	VARX, h = 1	VAR, h = 3	VARX, h = 3
1	0.1270	0.1271	0.2540	0.2537
2	0.1259	0.1250	0.2546	0.2521
3	0.1259	0.1266	0.2529	0.2548
7	0.1285	0.1301	0.2481	0.2544

Table 11 holds the 53 origins and VARX(2) structure fixed. Replacing differenced inflation with inflation in levels lowers RMSE from 0.1250 to 0.1246 at one month and from 0.2521 to 0.2479 at three months. The level specification thus has lower realized prediction errors in this sample, while Table 3 documents the mixed stationarity evidence underlying its sensitivity-only role. Inflation in levels represents the prevailing inflation state (for example, whether inflation is high), whereas differenced inflation represents the direction and magnitude of monthly change; the two carry different information. Empirically the level specification has slightly lower errors in this sample, but the mixed stationarity evidence leaves it on weaker statistical ground. Predictive loss and specification evidence are therefore two different criteria: the former answers which specification has smaller errors in this sample, the latter which is more likely to be stable, and the paper uses the latter to choose the primary input.

Table 11. CPI treatment and transformation sensitivity

Specification	RMSE, h=1	R^2_{OS}, h=1	RMSE, h=3	R^2_{OS}, h=3
Primary: Δinflation	0.1250	0.1611	0.2521	0.0955
Level inflation	0.1246	0.1659	0.2479	0.1256
Carry-forward then difference	0.1250	0.1611	0.2521	0.0956

6. Conclusion

The curve's own dynamics do carry genuine predictive value. All models are estimated on identical forecast origins. A parsimonious VAR in first differences, based on yields alone, outperforms the driftless random walk at both horizons. The one-month gain survives Holm adjustment. This result

DOI: 10.54254/2754-1169/2026.AB39126

holds for every lag order examined, so it is not an artefact of one specification. Macroeconomic information performs differently. Inflation, the policy rate, volatility and policy uncertainty reduce the errors only marginally once they are added. The reduction is statistically indistinguishable from zero. When the lag structure is extended, the increment becomes negative. So the additional contribution of these variables is weak and not robust. For short-horizon spread forecasts, a parsimonious yield system is already a demanding benchmark.

The distribution of the errors is informative. Almost all of the improvement over the benchmark is concentrated in months when the spread actually narrowed. When the spread widens, or when the curve is inverted, the augmented model is less accurate instead. The advantage of this class of models therefore depends on particular episodes, and it is not uniform across the evaluation window. One point deserves emphasis. The groupings are constructed from realised values in the target month, so what we describe is ex-post heterogeneity of pseudo-out-of-sample errors and nothing more. This pattern means that forecast difficulty varies with the direction of spread movements. It does not mean that a market state can be identified in advance, and it is not a state-dependent forecast.

The evidence on incremental conditional information is negative. Once the curve's own lags and the remaining macro variables are controlled for, lagged changes in the short rate contain no stable linear incremental information about the long rate. This test should not be interpreted causally. The finding is consistent with the out-of-sample evidence. With the joint dynamics of the curve included, the additional macro variables deliver neither a robust accuracy gain nor identifiable linear predictive content.

Three limitations should be noted. The baseline retains unexplained residual dynamics, and the cointegration evidence is sensitive to the levels lag order. Our claims therefore apply to this baseline only. All forecast origins fall after 2022, and they cover a single inversion and recovery, so the results need not extend to other periods. The data are revised rather than real-time, and the exercise is therefore pseudo-out-of-sample.

References

[1] Hansen, A. L. (2024). Predicting recessions using VIX–yield curve cycles. *International Journal of Forecasting, 40*(1), 409–422. https://doi.org/10.1016/j.ijforecast.2023.04.002

[2] Coe, P. J., & Vahey, S. P. (2025). Reassessing the predictive power of the yield spread for recessions in the United States. *Journal of Applied Econometrics, 40*(2), 231–236. https://doi.org/10.1002/jae.3106

[3] Choi, J., Ge, D., Kang, K. H., & Sohn, S. (2023). Yield spread selection in predicting recession probabilities. *Journal of Forecasting, 42*(7), 1772–1785. https://doi.org/10.1002/for.2980

[4] Leippold, M., & Matthys, F. (2022). Economic policy uncertainty and the yield curve. *Review of Finance, 26*(4), 751–797. https://doi.org/10.1093/rof/rfac031

[5] Rossi, B. (2021). Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them. *Journal of Economic Literature, 59*(4), 1135–1190. https://doi.org/10.1257/jel.20201479

[6] Demetrescu, M., Hanck, C., & Kruse-Becher, R. (2022). Robust inference under time-varying volatility: A real-time evaluation of professional forecasters. *Journal of Applied Econometrics, 37*(5), 1010–1030. https://doi.org/10.1002/jae.2906

[7] Doko Tchatoka, F., & Haque, Q. (2023). On bootstrapping tests of equal forecast accuracy for nested models. *Journal of Forecasting, 42*(7), 1844–1864. https://doi.org/10.1002/for.2987

[8] Huang, D., Jiang, F., Li, K., Tong, G., & Zhou, G. (2023). Are bond returns predictable with real-time macro data? *Journal of Econometrics, 237*(2), Article 105438. https://doi.org/10.1016/j.jeconom.2022.09.008

[9] Wan, R., Fulop, A., & Li, J. (2022). Real-time Bayesian learning and bond return predictability. *Journal of Econometrics, 230*(1), 114–130. https://doi.org/10.1016/j.jeconom.2020.04.052

[10] Fan, Y., Feng, G., Fulop, A., & Li, J. (2026). Real-time macro information and bond return predictability: A weighted group deep learning approach. *Journal of Financial Econometrics, 24*(5), Article nbag023.

DOI: 10.54254/2754-1169/2026.AB39126

https://doi.org/10.1093/jjfinec/nbag023
[11] Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond risk premiums with machine learning. *The Review of Financial Studies, 34*(2), 1046–1089. https://doi.org/10.1093/rfs/hhaa062
[12] Kang, H., & Ryu, D. (2026). Sentiment, uncertainty, and bond return predictability. *The Quarterly Review of Economics and Finance, 105*, Article 102083. https://doi.org/10.1016/j.qref.2025.102083
[13] Salachas, E., Kouretas, G. P., & Laopodis, N. T. (2024). The term structure of interest rates and economic activity: Evidence from the COVID-19 pandemic. *Journal of Forecasting, 43*(4), 1018–1041. https://doi.org/10.1002/for.3060
[14] Tvedt, J. (2025). A predictive term-spread model in the age of inflation targeting. *The North American Journal of Economics and Finance, 76*, Article 102364. https://doi.org/10.1016/j.najef.2025.102364
[15] Federal Reserve Bank of St. Louis. (n.d.). FRED: DGS2, DGS5, DGS10, CPIAUCSL, FEDFUNDS, VIXCLS, and USEPUINDXD [Data series]. https://fred.stlouisfed.org/
[16] Seip, K. L., & Zhang, D. (2021). The yield curve as a leading indicator: Accuracy and timing of a parsimonious forecasting model. *Forecasting, 3*(2), 421–436. https://doi.org/10.3390/forecast3020025