

Factor Covariance in Yield-Curve Risk Forecasts: Evidence from U.S. Treasury Yields

Fuli Yang* Shiyu Lin†

GARCH Institute, Harvard Square, Cambridge, MA 02138, USA

3 September 2026

Abstract

Yield-curve risk is often assembled in two steps: estimate separate volatilities for Nelson–Siegel factors, then map those volatilities back to individual maturities. The convenient diagonal map silently treats level, slope, and curvature innovations as conditionally uncorrelated. We show that this is not a harmless simplification for U.S. Treasury yields. Negative level–slope comovement makes diagonal forecasts too large at the front end, with a much smaller effect at long maturities. In a strictly out-of-sample exercise from 2008 to 2026, rolling CCC- and DCC-GARCH forecasts that retain factor covariance reduce QLIKE loss relative to diagonal GARCH at horizons of one, five, and twenty trading days. At two years, DCC lowers mean QLIKE by 0.77, 0.82, and 0.72, respectively, and reduces mean-squared error by 35–60 percent. The advantage narrows sharply with maturity. Two-sided interval exceedances provide the same diagnosis: the diagonal model covers too often because its intervals are excessively wide, whereas covariance-aware forecasts are much closer to nominal coverage. The improvement is not universal—it weakens at longer maturities and reverses for some twenty-day forecasts during the 2022–2023 tightening—but its maturity pattern follows directly from the Nelson–Siegel loadings. Factor covariance is therefore part of the forecast object, not a dispensable refinement to marginal volatility.

Keywords: Nelson–Siegel; yield-curve risk; multivariate GARCH; covariance forecasting; Treasury yields.

JEL classification: C32; E43; G12; G17.

1 Introduction

Many yield-curve risk systems are modular. A curve is compressed into level, slope, and curvature factors; a volatility model is fitted to each factor; and the resulting variances are mapped back to two-, five-, or ten-year yields. Each step is familiar. The problem lies in the join between them. Three well-estimated marginal variances do not form a covariance matrix, and a maturity is exposed to a linear combination of all three factors. Setting the missing off-diagonal terms to zero changes the risk forecast itself.

This paper asks whether that change matters outside the estimation sample. The question is deliberately narrower than choosing a globally optimal term-structure model. We hold the Nelson–Siegel representation and the marginal GARCH models fixed, and vary only the treatment of factor covariance. The benchmark discards covariance. Constant conditional correlation (CCC) restores a rolling correlation matrix. Dynamic conditional correlation (DCC) also lets that matrix move with incoming shocks. A backward-looking rolling-sample-covariance estimator provides a model-light comparison. We evaluate all four forecasts against subsequent changes in the *observed* Treasury constant-maturity yields, rather than against fitted Nelson–Siegel yields.

The answer is economically large and statistically visible at the front end. In the descriptive 2003–2026 sample, the level–slope innovation correlation is -0.84 . Because the Nelson–Siegel level and slope loadings are both positive, the omitted cross term is negative in the full quadratic form. Removing it raises the mapped variance. Holding marginal factor variances fixed, the diagonal-to-full conditional-variance ratio has a median of 2.90 at two years, 1.76 at five years, 1.37 at ten years, and 1.11 at thirty years.

That accounting result survives a harder forecasting test. We estimate models on rolling five-year windows, refit every 63 observations, and generate daily forecasts from January 2008 through August 2026. At two years, DCC-GARCH lowers average QLIKE loss relative to diagonal GARCH by 0.765 at one day, 0.823 at five days, and 0.724 at twenty days. The corresponding mean-squared-error ratios are 0.646, 0.537, and 0.396. CCC also improves on the diagonal benchmark, though less than DCC. The gains become smaller with maturity, just as the loading algebra predicts.

*Corresponding author. Email: dr@garch.ac.

†ORCID: [0009-0000-2471-3644](https://orcid.org/0009-0000-2471-3644).

Interval coverage makes the practical meaning transparent. For a nominal 5 percent two-sided exceedance rate, the one-day diagonal forecast produces only 0.8 percent exceedances at two years. It is not prudently calibrated; it is much too wide. CCC and DCC produce 5.1 and 4.9 percent. At twenty days the corresponding rates are 0.9, 3.9, and 5.2 percent. Covariance-aware models do not dominate every cell—five- and ten-year one-day intervals still undercover, and long-horizon improvements are imprecise at ten years—but they remove the diagonal model’s most pronounced front-end distortion.

We make three contributions. First, we isolate an exact and auditable channel through which factor dependence becomes maturity-specific risk error. Existing work establishes the usefulness of Nelson–Siegel factors and time-varying second moments; we focus on the final mapping from factor covariance to tenor risk. Second, we show that the mapping error has out-of-sample consequences under identical marginal volatility forecasts. This distinguishes a genuine covariance result from a comparison in which one model also changes the factor variances. Third, we document the boundary of the result. The benefit is strongest from two to five years, modest at ten to thirty years, and not uniformly positive during the 2022–2023 tightening. The paper therefore argues for preserving covariance, not for treating DCC as a universal winner.

The design also rules out two broader interpretations. This is not a bond-return predictability study: no expected-return model enters. It is not a portfolio-timing exercise: there is no allocation rule whose performance can hide a persistent risk bias. The empirical object is the conditional variance of yield changes at a stated tenor and horizon. Keeping that object narrow makes the comparison falsifiable.

The rest of the paper proceeds as follows. Section 2 links the question to term-structure and volatility forecasting. Section 3 defines the covariance map, competing forecasts, and real-time protocol. Section 4 reports the anatomy of covariance omission. Section 5 presents out-of-sample losses, interval coverage, and episode-level results. Section 6 discusses robustness and limitations. Section 7 concludes.

2 Related literature and empirical prediction

The level, slope, and curvature interpretation of Treasury yield movements predates the Nelson–Siegel forecasting literature and appears clearly in the principal-component evidence of [Litterman and Scheinkman \(1991\)](#). [Nelson and Siegel \(1987\)](#) supplied a parsimonious parametric loading system, and [Diebold and Li \(2006\)](#) gave its factors a dynamic forecasting interpretation. Subsequent work allowed the loading decay and factor dynamics to vary over time ([Koopman et al., 2010](#)), while [Christensen et al. \(2011\)](#) derived an affine arbitrage-free counterpart. These contributions make a low-dimensional risk representation natural, but none implies that factor innovations are independent. Our fitted-factor design is intentionally less structural: it isolates the second-moment mapping rather than asking the covariance model to repair pricing restrictions in the curve specification.

Term-structure models with stochastic volatility explicitly recognize time variation in second moments. [Hautsch and Yang \(2012\)](#) embeds stochastic volatility in a Nelson–Siegel system, while [Hautsch and Ou \(2012\)](#) studies interest-rate risk under stochastic factor volatility. [Shin and Zhong \(2017\)](#) show that realized-volatility information can improve bond-yield density forecasts. A parallel risk-management literature evaluates interest-rate covariance forecasts by the portfolio and interval decisions they support ([Ferreira and Lopez, 2005](#)). Our question is more operational and more tightly controlled. Once a researcher or risk manager has three factor volatility forecasts in hand, can the covariances be dropped when the forecasts are projected back to maturities? This question arises even in systems that do not attempt a fully joint state-space estimation.

The multivariate-GARCH literature provides several ways to retain dependence; [Bauwens et al. \(2006\)](#) survey the central specifications and their trade-offs. CCC imposes a fixed conditional correlation matrix while allowing marginal variances to change ([Bollerslev, 1990](#)). DCC permits parsimonious correlation dynamics ([Engle, 2002](#)). Its standard two-step estimator also has known interpretation and consistency limitations, motivating corrected formulations ([Aielli, 2013](#)). We therefore use DCC as a parsimonious forecasting competitor, not as a structural law for factor dependence. Comparing it with CCC separates two propositions: whether nonzero correlation matters at all, and whether allowing it to evolve adds further value.

Forecast evaluation follows two principles. First, the realized variance proxy can be noisy, so the loss function should not mechanically reward a model for matching that noise. The broad volatility-forecast comparisons of [Hansen and Lunde \(2005\)](#) demonstrate why out-of-sample ranking can depend on the chosen loss and proxy. QLIKE is robust to conditionally unbiased volatility proxies and is therefore the primary loss ([Patton, 2011](#)). For multivariate volatility forecasts, [Laurent et al. \(2012, 2013\)](#) likewise emphasize formal out-of-sample comparison and loss-function consistency. Mean-squared error is reported because scale errors are economically interpretable. Equal-loss inference follows the forecast-comparison logic of [Diebold and Mariano \(1995\)](#), with a HAC adjustment for overlapping horizons. Second, a risk model should also be judged by the events it is designed to cover. We report unconditional and independence diagnostics for two-sided forecast intervals following [Kupiec \(1995\)](#) and [Christoffersen \(1998\)](#).

The loading structure yields a directional prediction before any forecast is run. With positive level and slope loadings and negative level–slope covariance, deleting the cross term inflates tenor variance. Since the slope loading decreases with maturity, the inflation should be largest at short maturities and decay along the curve. If that mechanism is relevant out of sample, the loss advantage of covariance-aware forecasts should display the same maturity gradient.

3 Data and empirical design

3.1 Treasury yields and Nelson–Siegel factors

The data are daily Federal Reserve H.15 constant-maturity Treasury yields obtained through FRED. The cross section contains 3- and 6-month yields and 1-, 2-, 3-, 5-, 7-, 10-, 20-, and 30-year yields. The descriptive sample runs from 2 January 2003 through 27 August 2026. The out-of-sample evaluation begins on 2 January 2008, leaving approximately five years for the first estimation window. Yield units are percentage points.

For maturity τ in years, the Nelson–Siegel curve is

$$y_t(\tau) = \beta_{0t} + \beta_{1t} \frac{1 - e^{-\lambda\tau}}{\lambda\tau} + \beta_{2t} \left(\frac{1 - e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \right) = \mathbf{\Lambda}(\tau)' \boldsymbol{\beta}_t, \quad (1)$$

where $\lambda = 0.7308$ in the main specification. On every date with at least six observed maturities, we estimate $\boldsymbol{\beta}_t$ by cross-sectional ordinary least squares. The risk innovations are daily factor changes $\Delta\boldsymbol{\beta}_t$.

This fitted-factor approach is intentionally transparent. It avoids claiming that the three factors are observed without error, but it also keeps the covariance calculation reproducible from public data. The realized forecast target is not the fitted curve. For each reported tenor, we square the subsequent change in the corresponding raw CMT series. This choice asks the model to forecast an external observation rather than its own projection residual.

3.2 The covariance map

Let $\mathbf{H}_t = \text{Var}_t(\Delta\boldsymbol{\beta}_t)$ be the conditional factor covariance matrix. A one-day tenor-variance forecast is the quadratic form

$$V_t^{\text{full}}(\tau) = \mathbf{\Lambda}(\tau)' \mathbf{H}_t \mathbf{\Lambda}(\tau). \quad (2)$$

The diagonal forecast preserves the three conditional marginal variances while setting covariances to zero:

$$V_t^{\text{diag}}(\tau) = \sum_{i=0}^2 \Lambda_i(\tau)^2 H_{ii,t}. \quad (3)$$

Their difference is not an approximation residual of unknown origin. It is exactly

$$B_t(\tau) = V_t^{\text{diag}}(\tau) - V_t^{\text{full}}(\tau) = -2 \sum_{i < j} \Lambda_i(\tau) \Lambda_j(\tau) H_{ij,t}. \quad (4)$$

Equation (4) is useful because it separates magnitude from sign. A diagonal map is not inherently conservative. It is conservative in this sample only because the dominant omitted covariance is negative and multiplies positive loadings.

For $h \in \{1, 5, 20\}$ trading days, each GARCH model produces a path of factor covariance forecasts $\{\widehat{\mathbf{H}}_{t+k|t}\}_{k=1}^h$. We sum their tenor projections,

$$\widehat{V}_{t,h}(\tau) = \sum_{k=1}^h \mathbf{\Lambda}(\tau)' \widehat{\mathbf{H}}_{t+k|t} \mathbf{\Lambda}(\tau), \quad (5)$$

which treats forecasted daily innovations as serially uncorrelated. The realized proxy is $[y_{t+h}(\tau) - y_t(\tau)]^2$ from the raw CMT series.

3.3 Competing covariance forecasts

All GARCH specifications share zero-mean Student- t GARCH(1,1) marginal models for the three factor changes. The four forecasts differ only in how dependence is handled.

1. *Diagonal GARCH* sets the conditional correlation matrix to the identity.

2. *CCC-GARCH* combines the same marginal variances with the correlation matrix of standardized residuals in the current estimation window.
3. *DCC-GARCH* updates a correlation state $\mathbf{Q}_t$ according to

$$\mathbf{Q}_t = (1 - a - b)\bar{\mathbf{Q}} + a\mathbf{z}_{t-1}\mathbf{z}'_{t-1} + b\mathbf{Q}_{t-1}, \quad (6)$$

then standardizes $\mathbf{Q}_t$ into a correlation matrix. The parameters a and b are estimated by Gaussian correlation QML after fitting the marginal models.

4. *Rolling sample covariance* uses the sample covariance of factor changes over the previous 60 observations and scales its mapped daily variance by h .

The rolling-sample-covariance forecast is valuable precisely because it is not another elaborate parametric model. If it produces the same maturity ranking, the result is less likely to be an artifact of CCC or DCC likelihood choices. We reserve “realized volatility” for estimators constructed from higher-frequency observations; the present benchmark is simply a backward-looking daily sample covariance.

3.4 Real-time protocol and forecast evaluation

The first forecast uses the preceding 1,260 factor changes. Parameters are re-estimated every 63 observations on a rolling 1,260-observation window. Between refits, GARCH variances and DCC states are updated daily using only innovations already observed. No pre-2003 observations enter, and no full-sample correlation estimate enters the out-of-sample exercise. The protocol produces 4,667 one-day forecasts; slightly fewer observations are available at longer horizons.

The primary loss is

$$L_{t,h}^Q = \log \hat{V}_{t,h} + \frac{RV_{t,h}}{\hat{V}_{t,h}}, \quad (7)$$

where $RV_{t,h}$ is the squared raw-yield change. Lower QLIKE is better. We also report MSE and the ratio of each model’s MSE to diagonal GARCH. For loss differences, heteroskedasticity-and-autocorrelation-consistent standard errors use a Bartlett kernel with lag $\max(5, 2h)$ (Diebold and Mariano, 1995; Newey and West, 1987). This adjustment is essential because the five- and twenty-day forecast outcomes overlap.

For coverage, the models define symmetric Gaussian intervals with half-width $z_{0.975}\sqrt{\hat{V}_{t,h}}$. The distributional assumption is deliberately held fixed across models so that coverage differences reflect variance forecasts rather than different tail quantiles. Multi-day coverage tests use nonoverlapping observations. We report the observed exceedance frequency and p -values from unconditional-coverage and exceedance-independence tests.

4 The anatomy of covariance omission

The full-sample CCC correlations are

$$\text{Corr}(\Delta\beta_0, \Delta\beta_1) = -0.84, \quad \text{Corr}(\Delta\beta_0, \Delta\beta_2) = -0.02, \quad \text{Corr}(\Delta\beta_1, \Delta\beta_2) = -0.22.$$

The level-slope term is therefore the natural place to look for the maturity gradient.

Table 1: The measurement effect of omitting factor covariance

Tenor	Mean V^{full}	Mean V^{diag}	Med($V^{\text{diag}}/V^{\text{full}}$) [95% interval]	Rolling 60d median
2 years	0.00283	0.00752	2.897 [2.784, 3.006]	3.247
5 years	0.00320	0.00554	1.760 [1.736, 1.782]	1.785
10 years	0.00294	0.00407	1.375 [1.367, 1.381]	1.362
30 years	0.00313	0.00348	1.109 [1.107, 1.111]	1.101

Notes: 5,917 daily factor changes, 3 January 2003–27 August 2026. The CCC- t comparison holds marginal conditional variances fixed. Intervals use 5,000 circular moving-block resamples of length 60, conditional on the fitted full-sample model. The last column uses only the preceding 60 factor changes. Variance units are squared percentage points.

Table 1 shows that the diagonal map is not a small perturbation. At two years, it produces almost three times the variance of the full map on the median day. The ratio falls monotonically and is close to one by thirty years. The backward-looking estimator yields essentially the same ordering, with an even larger median ratio at two years.

The mechanism is also stable through time. Figure 1 plots the level–slope correlation estimated in each rolling out-of-sample window. It remains negative throughout and ranges from approximately -0.97 to -0.63 . This persistence explains why the diagonal error is not confined to a single policy episode.

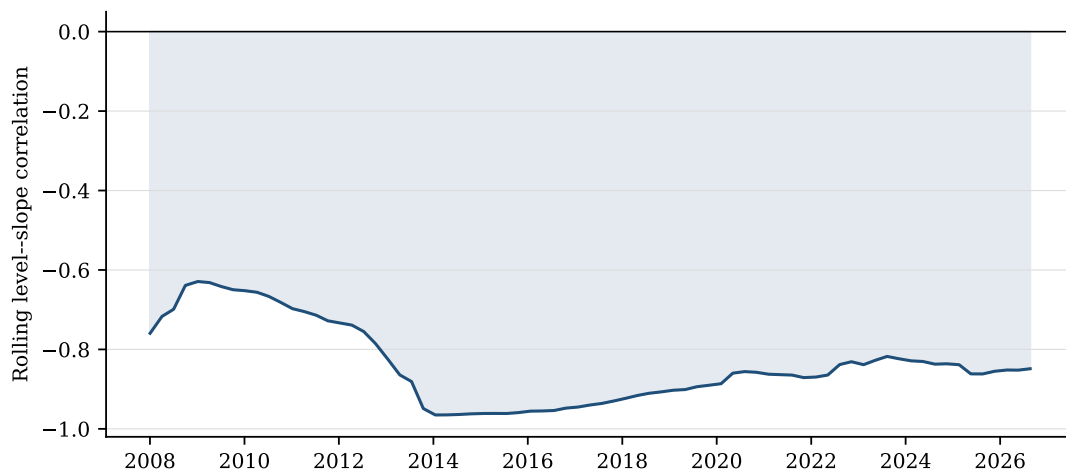


Figure 1: Rolling level–slope correlation. Each point uses standardized marginal-GARCH residuals from the preceding 1,260 observations at an out-of-sample refit date.

Decomposing Eq. (4) confirms that level–slope covariance accounts for about 79 percent of the mean two-year diagonal-minus-full difference. The level–curvature and slope–curvature terms are smaller. The absolute level–slope contribution declines from 0.00372 squared percentage points at two years to 0.00032 at thirty years. This is the algebraic source of the forecast prediction tested next.

5 Out-of-sample results

5.1 Forecast losses

Table 2 is the main result. Panel A reports each covariance-aware model’s mean QLIKE difference relative to diagonal GARCH; negative entries indicate improvement. Panel B reports MSE ratios, with values below one favoring the covariance-aware model.

Table 2: Out-of-sample variance-forecast performance, 2008–2026

h	Tenor	Panel A: QLIKE difference vs. diagonal			Panel B: MSE ratio vs. diagonal		
		CCC	DCC	Rolling 60d	CCC	DCC	Rolling 60d
1	2y	−0.528***	−0.765***	−0.743***	0.678	0.646	0.670
1	5y	−0.096***	−0.145***	−0.102***	0.897	0.872	0.886
1	10y	−0.021**	−0.035***	−0.018	0.973	0.960	0.967
1	30y	−0.014***	−0.017***	−0.010	0.984	0.974	0.988
5	2y	−0.554***	−0.823***	−0.879***	0.553	0.537	0.510
5	5y	−0.110***	−0.170***	−0.159***	0.784	0.742	0.720
5	10y	−0.031*	−0.052***	−0.052**	0.933	0.899	0.892
5	30y	−0.021***	−0.025***	−0.028***	0.969	0.952	0.963
20	2y	−0.447***	−0.724***	−0.806***	0.423	0.396	0.346
20	5y	−0.104*	−0.154**	−0.142*	0.747	0.682	0.630
20	10y	−0.024	−0.051	−0.051	0.941	0.890	0.867
20	30y	−0.023**	−0.028***	−0.039**	0.975	0.952	0.963

Notes: Forecasts use rolling 1,260-observation windows and quarterly parameter refits. Realizations are squared changes in raw CMT yields. Negative QLIKE differences and MSE ratios below one favor the column model. HAC inference uses lag $\max(5, 2h)$. *, **, and *** denote 10, 5, and 1 percent significance for equal mean QLIKE loss.

Three features matter. First, retaining covariance improves the two-year forecast in every horizon and loss function. The DCC MSE reduction grows from 35 percent at one day to 60 percent at twenty days. Second, DCC typically improves on CCC, especially at the front end. Correlation dynamics add information beyond a nonzero rolling correlation matrix. Third, the maturity gradient is steep. Improvements remain detectable at five years, are much smaller at ten years, and are economically modest at thirty years even when statistically significant.

The rolling-sample-covariance benchmark is competitive. It records the lowest two-year MSE at five and twenty days and the largest twenty-day QLIKE improvement. That result is useful rather than embarrassing: the paper’s claim concerns preserving covariance, not the superiority of a particular parametric estimator. Both the parametric and model-light routes reject the diagonal restriction.

Figure 2 plots the running mean DCC-minus-diagonal QLIKE difference for twenty-day forecasts. The two-year line stays well below zero across most of the evaluation. The long-maturity lines lie much closer to zero. The graph also reveals a deterioration around the recent tightening cycle, which motivates the episode analysis below.

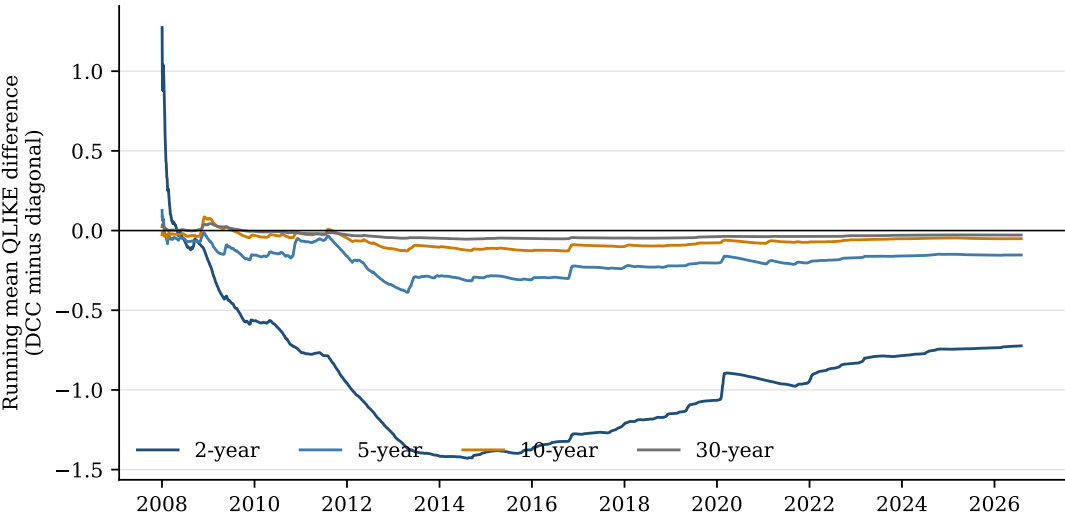


Figure 2: Running mean QLIKE difference for twenty-day forecasts. Negative values favor DCC-GARCH. Outcomes overlap; statistical inference is reported in Table 2 using HAC standard errors.

5.2 Forecast-interval coverage

Average losses do not show whether a model is systematically too wide or too narrow. Table 3 reports two-sided exceedance rates for nominal 95 percent intervals.

Table 3: Two-sided 95 percent forecast-interval exceedance rates

Horizon	Model	2 years	5 years	10 years	30 years
1 day	Diagonal GARCH	0.8***	2.0***	3.0***	3.4***
	CCC-GARCH	5.1	6.7***	6.2***	4.3**
	DCC-GARCH	4.9	6.9***	5.8***	4.2***
20 days	Diagonal GARCH	0.9***	1.7***	3.0	1.7***
	CCC-GARCH	3.9	7.3	6.4	3.0
	DCC-GARCH	5.2	6.9	7.3	3.0

Notes: Entries are percentages. The nominal exceedance rate is 5 percent. Twenty-day tests use every twentieth forecast to avoid overlapping coverage events. Stars denote rejection of correct unconditional coverage using the Kupiec test: ** 5 percent and *** 1 percent. The same Gaussian critical value is used for all models.

The diagonal model’s front-end intervals are far too wide. Its one-day two-year interval has a mean half-width of 0.165 percentage points, compared with 0.089 for DCC, and records fewer than one fifth of the intended exceedances. At twenty days the mean half-width is 0.745 percentage points under the diagonal model and 0.405 under DCC. DCC restores the two-year exceedance rate almost exactly to nominal at both horizons.

Coverage elsewhere is not perfect. At one day, CCC and DCC intervals are somewhat too narrow at five and ten years, while thirty-year intervals are slightly wide. At twenty days, sample sizes are much smaller after nonoverlapping selection and the tests have correspondingly low power. The important point is narrower: treating negative factor covariance as zero creates severe and predictable front-end overcoverage.

5.3 Episode dependence

The full-sample averages could conceal a single crisis payoff. We therefore report twenty-day DCC-minus-diagonal QLIKE differences for the global financial crisis, the 2020 pandemic, the 2022–2023 tightening, and all other dates.

Table 4: Twenty-day DCC QLIKE difference across episodes

Episode	2 years	5 years	10 years	30 years
Global financial crisis	−0.728	−0.145	0.034	0.026
Pandemic	−1.498	−0.857	−0.360	−0.089
2022–2023 tightening	0.330	0.103	0.092	0.032
Other dates	−0.820	−0.148	−0.057	−0.036

Notes: Entries are mean QLIKE for DCC-GARCH minus mean QLIKE for diagonal GARCH. Negative values favor DCC. The global financial crisis is September 2008–June 2009; the pandemic period is March–December 2020; tightening is January 2022–December 2023.

The covariance gain is not merely a crisis effect. At two years it remains large on dates outside the three named episodes. It is especially strong during the pandemic, when factor dependence moved quickly and the dynamic correlation model had room to adjust. During the 2022–2023 tightening, however, DCC loses to the diagonal model at every reported maturity for twenty-day QLIKE. This reversal does not overturn the full-sample result, but it warns against interpreting a historically negative level–slope covariance as a fixed structural hedge. Rapid policy repricing can alter both the persistence of correlation and the serial dependence ignored in Eq. (5).

6 Robustness, scope, and limitations

The original covariance measurement is stable to three checks. A purely backward-looking 60-day covariance produces the same sign and maturity ranking. Splitting the descriptive sample into 2003–2012, 2013–2019, and 2020–2026 yields two-year median diagonal/full ratios of 2.86, 3.17, and 2.53. Re-estimating factors and volatility models over $\lambda \in \{0.40, 0.60, 0.7308, 1.00\}$ gives two-year ratios between 2.69 and 3.29, while thirty-year ratios remain between 1.09 and 1.17. These exercises show that the front-end concentration is not created by one decay calibration.

The out-of-sample design is deliberately harder than the descriptive calculation. All correlation estimates are rolling; the raw Treasury yield, not the NS fitted value, supplies the realization; and the model is evaluated at three pre-specified horizons. The comparable performance of the rolling 60-day sample covariance also shows that the main finding is not dependent on a particular DCC optimizer.

Several limitations remain. First, Eq. (5) sums daily conditional variances and excludes autocovariances of factor changes. That approximation is common but can matter at the twenty-day horizon, particularly during rapid tightening. A joint model for factor means and covariance would address it, at the cost of changing more than the off-diagonal risk object. Second, the Student- t assumption enters the marginal GARCH estimates, but interval coverage is compared using a common Gaussian critical value. This choice isolates variance calibration; it is not a claim that Gaussian tails are adequate for Treasury changes. Third, CMT yields are interpolated benchmark series rather than tradable zero-coupon yields. Replication with zero-coupon curves and intraday realized measures would sharpen the economic interpretation.

Finally, a larger variance estimate is not automatically a safer risk estimate. The diagonal model appears conservative because it overstates front-end variance in this sample. Yet the cost is distorted relative risk across maturities and intervals that cover far more often than intended. A risk manager can choose an explicit capital buffer after obtaining a calibrated forecast. Embedding a maturity-dependent buffer accidentally by deleting covariance is harder to defend.

7 Conclusion

The central empirical fact in this paper is simple: Nelson–Siegel factor covariance survives the map back to maturity. For U.S. Treasury yields, strongly negative level–slope dependence subtracts substantial variance at the front end. A diagonal map deletes that offset, inflating two-year risk far more than thirty-year risk.

The out-of-sample evidence shows that this is more than descriptive accounting. Rolling CCC, DCC, and sample-covariance forecasts generally reduce QLIKE and MSE relative to diagonal GARCH, with the largest

gains at two years. DCC also converts severely overwide front-end forecast intervals into intervals close to nominal coverage. The advantage weakens along the curve and is not uniform across policy episodes, especially at twenty-day horizons during the recent tightening.

The practical conclusion is therefore precise. Separate factor volatility models can remain useful, but their outputs should not be mapped to tenor risk through an identity correlation matrix without testing the implied cross terms. Covariance is not an optional decoration on three marginal forecasts. It is part of the maturity-level risk forecast they are meant to produce.

Data and code availability

The Federal Reserve H.15 constant-maturity yield series are publicly available through FRED. Series identifiers, raw data, estimation code, forecast outputs, and figure code are included in the replication package accompanying this manuscript.

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies

During preparation of this manuscript, the authors used OpenAI's ChatGPT for language editing and code assistance. The authors reviewed and edited all output and take full responsibility for the content.

A Implementation details

At each refit date, the three marginal models are estimated on the previous 1,260 observations after multiplying factor changes by 100 for numerical stability. Conditional variances are transformed back to squared percentage points before tenor mapping. CCC correlations use standardized residuals from the same rolling window. DCC parameters satisfy $a \geq 0$, $b \geq 0$, and $a + b < 0.999$. Across 75 refits, the median estimates are $a = 0.066$ and $b = 0.927$; the median rolling CCC level-slope correlation is -0.858 .

Within each 63-observation block, parameters are fixed but states are not. Before forecasting observation t , the marginal variance recursion uses the shock observed at $t - 1$, and the DCC recursion uses standardized residuals through $t - 1$. After recording all horizon forecasts for t , the realized factor change updates the states used at $t + 1$. This ordering prevents a same-day shock from entering its own forecast.

Sampling intervals in Table 1 are descriptive circular moving-block intervals conditional on the fitted full-sample model; they do not incorporate parameter-estimation uncertainty. By contrast, significance in Table 2 is based on the out-of-sample loss-difference series. The two forms of uncertainty answer different questions and are not combined.

References

- Aielli, G.P., 2013. Dynamic conditional correlation: On properties and estimation. *Journal of Business & Economic Statistics* 31, 282–299. <https://doi.org/10.1080/07350015.2013.771027>.
- Bauwens, L., Laurent, S., Rombouts, J.V.K., 2006. Multivariate GARCH models: A survey. *Journal of Applied Econometrics* 21, 79–109. <https://doi.org/10.1002/jae.842>.
- Bollerslev, T., 1990. Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Review of Economics and Statistics* 72, 498–505. <https://doi.org/10.2307/2109358>.

- Christensen, J.H.E., Diebold, F.X., Rudebusch, G.D., 2011. The affine arbitrage-free class of Nelson–Siegel term structure models. *Journal of Econometrics* 164, 4–20. <https://doi.org/10.1016/j.jeconom.2011.02.011>.
- Christoffersen, P.F., 1998. Evaluating interval forecasts. *International Economic Review* 39, 841–862. <https://doi.org/10.2307/2527341>.
- Diebold, F.X., Li, C., 2006. Forecasting the term structure of government bond yields. *Journal of Econometrics* 130, 337–364. <https://doi.org/10.1016/j.jeconom.2005.03.005>.
- Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13, 253–263. <https://doi.org/10.1080/07350015.1995.10524599>.
- Engle, R.F., 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics* 20, 339–350. <https://doi.org/10.1198/073500102288618487>.
- Ferreira, M.A., Lopez, J.A., 2005. Evaluating interest rate covariance models within a value-at-risk framework. *Journal of Financial Econometrics* 3, 126–168. <https://doi.org/10.1093/jjfinec/nbi005>.
- Hansen, P.R., Lunde, A., 2005. A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics* 20, 873–889. <https://doi.org/10.1002/jae.800>.
- Hautsch, N., Ou, Y., 2012. Analyzing interest rate risk: Stochastic volatility in the term structure of government bond yields. *Journal of Banking & Finance* 36, 2988–3007. <https://doi.org/10.1016/j.jbankfin.2012.06.020>.
- Hautsch, N., Yang, F., 2012. Bayesian inference in a stochastic volatility Nelson–Siegel model. *Computational Statistics & Data Analysis* 56, 3774–3792. <https://doi.org/10.1016/j.csda.2010.07.003>.
- Koopman, S.J., Mallee, M.I.P., van der Wel, M., 2010. Analyzing the term structure of interest rates using the dynamic Nelson–Siegel model with time-varying parameters. *Journal of Business & Economic Statistics* 28, 329–343. <https://doi.org/10.1198/jbes.2009.07295>.
- Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives* 3, 73–84. <https://doi.org/10.3905/jod.1995.407942>.
- Laurent, S., Rombouts, J.V.K., Violante, F., 2012. On the forecasting accuracy of multivariate GARCH models. *Journal of Applied Econometrics* 27, 934–955. <https://doi.org/10.1002/jae.1248>.
- Laurent, S., Rombouts, J.V.K., Violante, F., 2013. On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics* 173, 1–10. <https://doi.org/10.1016/j.jeconom.2012.08.004>.
- Litterman, R., Scheinkman, J., 1991. Common factors affecting bond returns. *Journal of Fixed Income* 1, 54–61. <https://doi.org/10.3905/jfi.1991.692347>.
- Nelson, C.R., Siegel, A.F., 1987. Parsimonious modeling of yield curves. *Journal of Business* 60, 473–489. <https://doi.org/10.1086/296409>.
- Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–708. <https://doi.org/10.2307/1913610>.
- Patton, A.J., 2011. Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics* 160, 246–256. <https://doi.org/10.1016/j.jeconom.2010.03.034>.
- Shin, M., Zhong, M., 2017. Does realized volatility help bond yield density prediction? *International Journal of Forecasting* 33, 373–389. <https://doi.org/10.1016/j.ijforecast.2016.11.003>.