Modeling and Forecasting the Co-Movement of International Yield Curve Drivers

Stefan Mittnik and Maria Sprincenatu

Munich, Germany

Abstract

We propose a Full State-Space Model (FSSM) to jointly model and forecast the U.S. and German yield curves. The model embeds the empirically observed dynamic features of both curves within a unified VAR-VEC state-space representation. The FSSM nests country-specific versions, allowing us to quantify the marginal value of foreign information for domestic prediction.

Estimation is performed on an identification sample (1999-2018), while forecast evaluation uses a long out-of-sample period (1999-2025) covering the inflation shock, global tightening cycle, and post-pandemic normalization of interest rates. We compare the FSSM with the Random Walk, domestic and global Diebold-Li models, and Germany-only and US-only variants of the FSSM (FSSM-DE and FSSM-US).

In point forecasts, the Random Walk remains the strongest benchmark, and the FSSM only outperforms the Diebold-Li benchmarks at medium and long horizons. In contrast, the FSSM delivers superior directional accuracy and the best density forecasts across all metrics. Its densities are sharper, better calibrated, and more coherent across maturities and horizons. The Germany-only and U.S.-only versions underperform the full cross-country FSSM, showing that foreign

information substantially improves domestic yield-curve forecasts—consistent with global rate spillovers.

Overall, the results highlight the importance of structural cross-country linkages for reliable yield-curve prediction and demonstrate the forecasting gains from modeling the joint dynamics of global term structures.
Keywords: Term Structure, International Yield Curves, State-Space Models, Point Forecasting, Directional Evaluation, Density Forecasting, Cross-Country Spillovers

---

1. Introduction

Modeling and forecasting the yield curve has long been a central topic among monetary policymakers, academics, and fixed-income practitioners. In increasingly integrated global capital markets, yield curves across major currency areas exhibit strong contemporaneous and lagged co-movement, giving rise to rich cross-country spillovers and shared policy influences. These linkages imply that central banks, international investors, and risk managers require forecasting frameworks capable not only of capturing domestic yield dynamics but also of modeling the joint evolution of global term structures.

The Diebold-Li "yields-only" model (Diebold et al. (2006)) has become a popular benchmark, especially after the 2008 Financial Crisis, owing to its parsimony, ease of estimation, strong long-horizon forecastability, and straightforward extension to a global setting (Diebold & Li (2006); Diebold et al. (2008); Diebold & Rudebusch (2013)). Yet, despite its empirical success, the model abstracts from key in-sample properties of yields and yield curve factors. In particular, yield curves are well known to exhibit highly persistent, potentially non-stationary

[I(1)] behavior, suggesting the presence of cointegration relationships among yields both within and across countries. Standard implementations of Diebold-Li ignore these features and impose strong covariance restrictions that rule out contemporaneous and non-contemporaneous global dependencies (Belke & Gros (2005); Stock & Watson (2005); Rey (2016)).

A further limitation is that the global yield curve literature has paid little attention to structural breaks, despite the availability of well-established econometric tools (Bai et al. (1998); Bai et al. (2000); Hansen (2003); Qu & Perron (2007); Commandeur & Koopman (2007); Commandeur et al. (2011); Durbin & Koopman (2012)). Popular global term structure models such as Diebold et al. (2008) assume stable parameters and fit global factors with a stationary VAR(1) process. This assumption may be problematic, as structural breaks in monetary policy regimes—notably around the 2008 Financial Crisis—affect central bank predictability (Hanspeter (2004); ECB (2011a); ECB (2011b); ECB (2011c); Wyplosz (2013); De La Dehesa (2013); Rodriguez et al. (2014); Verhelst (2014); Fed (2018)) and can meaningfully alter the dynamics of the yield curve. Understanding how such structural changes propagate across countries remains an open empirical question.

While there exists an extensive literature on modeling the term structure, studies on forecasting—especially in a multi-country setting—are relatively limited. Arbitrage-free (Hull & White (1990); Heath et al. (1992)) and affine models (Vasicek (1977); Cox et al. (1977); Duffie & Kan (1996)) typically perform poorly out-of-sample (Duffee (2002)). Consequently, domestic (Nelson & Siegel (1987); Litzenberger et al. (1995); Balduzzi et al. (1996); Chen (1996); Bliss (1997a); Bliss (1997b); Andersen & Lund (1997); Dai & Singleton (2000); De Jong & Santa-Clara

(1999); Jong (2000); Brandt & Yaron (2003); Duffee (2002)) and global factor models (Diebold et al. (2008); Jotikasthira et al. (2015)) have become the dominant forecasting tools, with the Diebold-Li model frequently outperforming random walk, slope regression, and AR benchmarks. Yet, only few studies—most prominently Diebold et al. (2008)—explicitly model the joint evolution of multiple country yield curves, and even these models impose strong dynamic and covariance restrictions.

Moreover, existing models rarely explore forecast performance beyond point predictions. Modern risk management, asset allocation, and policy analysis increasingly rely on directional accuracy, probability forecasts, and full predictive distributions rather than point estimates alone. Little evidence exists on how domestic, global, and structural-state-space yield curve models compare in terms of interval coverage, directional accuracy, or density forecast performance.

This paper addresses these gaps by developing a family of *Full State-Space Models* (FSSM) for jointly forecasting the co-movement of the U.S. and German yield curves. The models preserve the empirical dynamic properties of the underlying yield curve factors, allow for cointegration, incorporate contemporaneous and lagged cross-country transmission, and flexibly accommodate structural breaks in the data. Alongside the full U.S.-Germany FSSM, we also construct two restricted variants: FSSM-US (forecasting the U.S. term structure with zero restrictions on German parameters) and FSSM-DE (forecasting Germany with zero restrictions on U.S. parameters). These variants allow us to quantify the value of cross-country information, particularly whether U.S. dynamics help forecast German yields.

Using an extended out-of-sample period (2018-2025), we evaluate the models using point forecasts, directional accuracy, and full predictive densities (predictive intervals, Energy Scores, and Joint Log Scores). We find that while the

Random Walk remains difficult to beat in point RMSE, the FSSM consistently outperforms Domestic and Global Diebold-Li benchmarks at medium and long horizons and delivers markedly superior directional accuracy and density forecasts than all benchmarks. Furthermore, the FSSM-DE performs worse than the FSSM, showing that incorporating U.S. information materially improves German yield forecasts—consistent with the view that Fed policy leads global yield curve dynamics, with the ECB often reacting with some delay. Similarly, deterioration in FSSM-US forecasts suggests that German yield movements embed information about forthcoming U.S. rate dynamics, in line with cross-market linkages in the global interest-rate cycle.

We proceed as follows. Section 2 analyzes the dynamic properties and structural breaks in the international yield curve drivers. Section 3 develops the structured state-space framework, its estimation, and its extensions to account for structural changes. Section 4 presents the out-of-sample point forecasting results and Section 5 the evaluation of directional accuracy. Sections 6 and 7 are dedicated to the density forecasting results with bootstrapping and Monte Carlo methods, respectively. Section 8 concludes.

2. Data Description, Vizualization, and Dynamic Properties

Our modeling approach consists in first conducting a comprehensive study of the dynamic properties of international yield curve drivers. To this end, we employ methods such as unit root tests, cross-correlation analysis, Granger causality, cointegration tests, and impulse response analysis to derive data-generation processes that best capture the co-movement of the U.S. and German yield curve drivers. Additionally, we investigate the presence of outliers and structural breaks

in these processes.

2.1. Data Description and Visualization

For the development of our data-driven state-space models, we use actively traded government bond yields[1] for the U.S. and Germany. The *identification sample* used to study dynamic properties and structural breaks runs from 1999:01 to 2018:01. The yield data are sampled monthly (229 monthly observations) and the cross sections span short-, medium-, and long-term maturities, i.e., 6-month, 1-year, 2-year, 3-year, 5-year, 7-year, and 10-year maturities. In the forecasting section, we embed this identification window into a longer *extended sample* (1999:01 to 2025:06) that allows us to evaluate point, directional, and density forecasts over a prolonged out-of-sample period.

In Figure 1, we show the U.S. and German government bond yield curves in levels and first differences. Complex movements in levels, slopes, and curvatures are readily apparent. The yields in levels appear highly persistent and exhibit a downward drift over the sample, which is more pronounced for German yields. The high persistence of yields in levels suggests nonstationary dynamics, whereas the first differences of the yields display more stationary behavior. With respect to volatility, the U.S. first differences exhibit higher variability compared to Germany. The bottom panels of Figure 1 display the three-dimensional surface of the German-U.S. yield spreads, both in levels and first differences. The spreads themselves appear somewhat less persistent than the underlying yields, which is consistent

[1]The data have been retrieved from the Federal Reserve Board Federal Reserve Board (FRB): Download Program. and Deutsche BundesbankDeutsche Bundesbank: Time series databases. databases

with cross-country co-movement and partial mean reversion of yield differentials.

It is easy to spot the period of the 2007-2008 Financial Crisis, whose onset is marked by a drastic decline in yields and the subsequent low interest rate environment. In the U.S., the low-rate regime started soon after December 2008, when the Federal Reserve System (the Fed) reduced the federal funds rate to effectively zero and initiated the first round of quantitative easing. In Germany, the low-rate environment began between October 2008 and May 2009, when the European Central Bank (ECB) reduced its key policy rates and introduced "enhanced credit support" to the banking sector. During this period, German yields entered negative territory at several maturities. These drastic central-bank interventions introduced substantial noise and regime changes into the term structure dynamics. The noise is visible in the first-difference surfaces, which show large movements around mid-2008 and subsequent episodes of elevated volatility.

Descriptive statistics[2] of the yield data, including sample autocorrelations, confirm the high persistence of yields in levels for both countries.

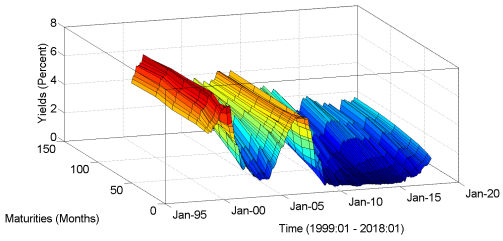
Because the level, slope, and curvature (Nelson & Siegel (1987); Litterman & Scheinkman (1991); Dai & Singleton (2000); Diebold & Li (2006)) account for most of the variation in the shape of the yield curves, we next analyze the level, slope, and curvature of the U.S. and German term structures. We estimate these unobservable factors via the dynamic Nelson-Siegel model (Diebold & Li (2006)):

$$y_{it}(\tau) = l_{it} + s_{it}\left(\frac{1-e^{-\lambda_{it}\tau}}{\lambda_{it}\tau}\right) + c_{it}\left(\frac{1-e^{-\lambda_{it}\tau}}{\lambda_{it}\tau} - e^{-\lambda_{it}\tau}\right) + \nu_{it}(\tau), \quad (1)$$

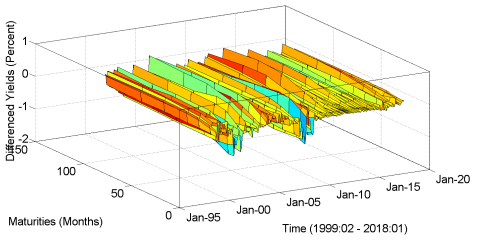
where l_{it}, s_{it}, and c_{it} denote the (U.S. or German, $i = US, DE$) country-specific

[2]Available from the authors on demand.

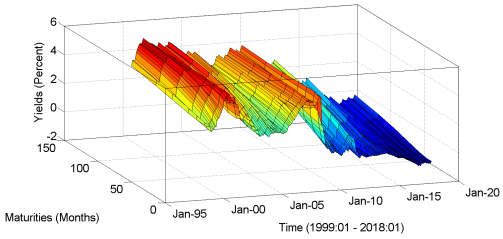
Figure 1: Yield curves over space and time. (Notes: Monthly data, 1999:01-2018:01, for 6-month, 1-year, 2-year, 3-year, 5-year, 7-year, and 10-year maturities).



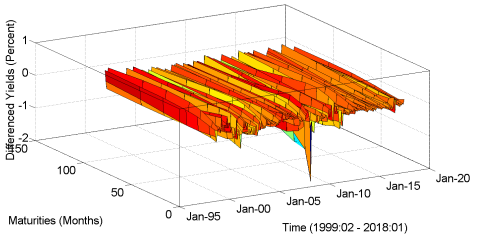
(a) U.S. yields, in levels.



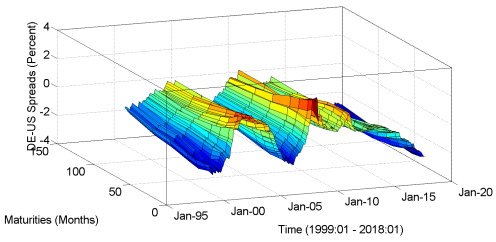
(b) U.S. yields, in first differences.



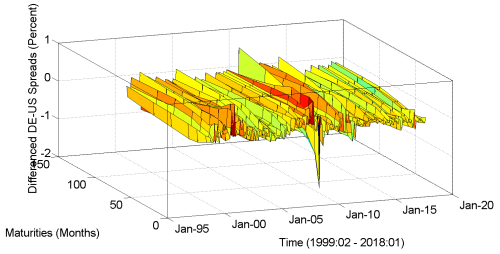
(c) German yields, in levels.



(d) German yields, in first differences.



(e) German-U.S. yield spreads, in levels.

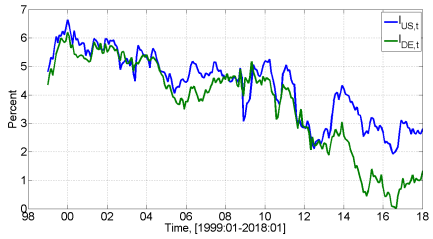


(f) German-U.S. yield spreads, in first differences.

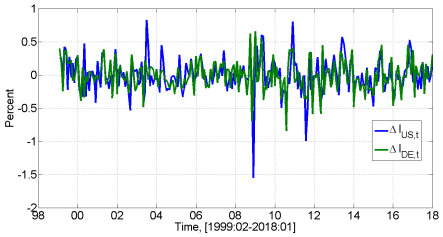
level, slope, and curvature factors. The exponential components represent the Nelson-Siegel loading structure, which controls how the three factors affect yields at different maturities.

In Figures 2, 3, and 4, we show the U.S. and German level, slope, and curvature factors (in levels and first differences).

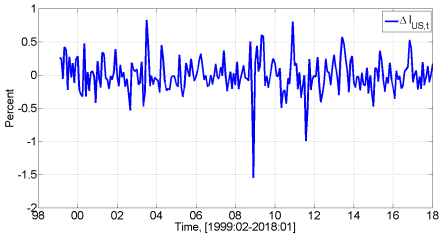
Figure 2: (Dynamic Nelson-Siegel) estimated country factors: U.S. and German levels, 1999:01-2018:01.



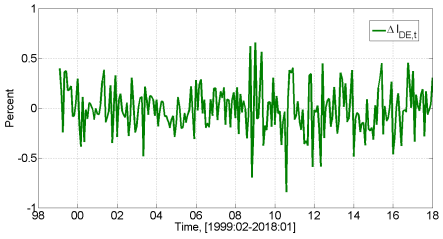
(a) U.S. and German levels, in levels.



(b) U.S. and German levels, in first differences.



(c) U.S. level, in first differences.

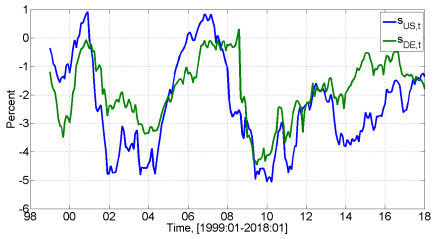


(d) German level, in first differences.

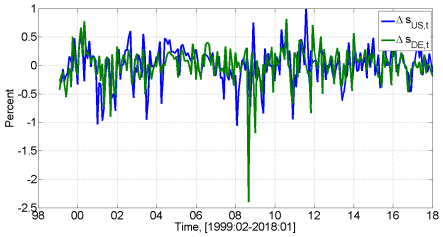
These plots reveal a high degree of commonality in the movements of the factors: U.S. and German factors tend to move together and follow broadly similar dynamics. The levels exhibit a decreasing trend over time that points to nonstationarity. From around the end of 2012 to the end of the identification sample, their behavior diverges: the U.S. level begins an upward trend, whereas

the German level remains much lower and even approaches zero around early 2017. This suggests persistent cross-country differentials in the long end of the curve. The first differences of the levels show more contained volatility, consistent with approximate stationarity, together with episodes of volatility clustering between 2009 and 2013, where large changes tend to follow large changes and small changes follow small changes.

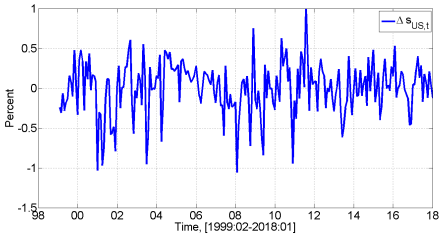
Figure 3: (Dynamic Nelson-Siegel) estimated country factors: U.S. and German slopes, 1999:01-2018:01.



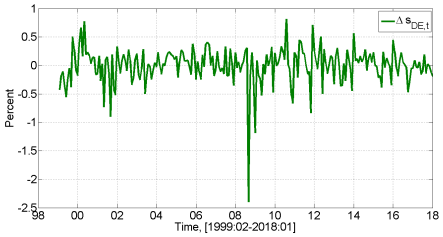
(a) U.S. and German slopes, in levels.



(b) U.S. and German slopes, in first differences.



(c) U.S. slope, in first differences.



(d) German slope, in first differences.

Similar observations apply to the U.S. and German slopes and curvatures, shown in Figures 3 and 4. All three factors exhibit strong intra- and inter-country co-movement, with occasional periods in which one country's curve steepens or flattens more markedly than the other.

An intra- and inter-country correlation analysis[3] shows positive correlations, significantly different from zero, for the U.S. level and curvature and for the U.S. slope and curvature. The same analysis for Germany shows that all German yield curve drivers have correlations significantly different from zero. Inter-country factor correlations are significantly different from zero for almost all pairs of mixed U.S. and German yield curve drivers, except for $s_{US,t} - l_{DE,t}$ and $c_{US,t} - l_{DE,t}$. Because some heteroskedasticity is visible in the matrix plot of inter-country factor correlation, we investigate the presence of volatility clustering in the residuals of differenced country factors with Engle's ARCH test. The null hypothesis of no conditional heteroskedasticity is rejected only for $\Delta c_{DE,t}$.

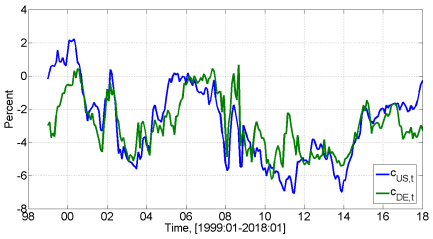
2.2. Dynamic Properties

To gain a deeper understanding of the dynamic properties of the U.S. and German yield curve drivers, we follow a structured workflow. First, we assess the nonstationary behavior of the yield curve variables via unit root tests. At this stage, we find that yields, yield spreads, U.S. and German country levels, and U.S. and German country curvatures are nonstationary variables, integrated of order one. This finding sits uncomfortably with the common theoretical argument that nominal bond yields should be stationary around a bounded path, but it is consistent with the persistent low-rate environment and regime changes observed since the late 1990s.

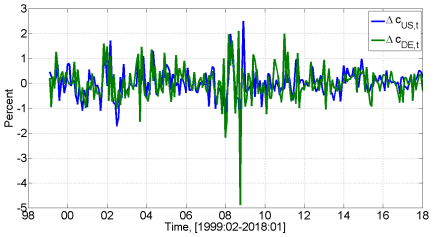
Next, we run a cross-correlation analysis with the objective of exploring more deeply the commonality in movements and the presence of lead-lag structure

[3]The plots of this analysis are available from the authors on demand.

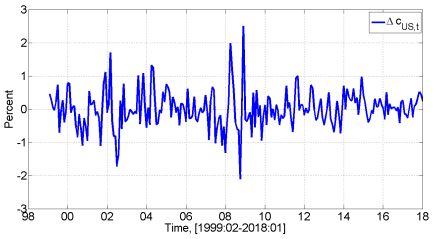
Figure 4: (Dynamic Nelson-Siegel) estimated country factors: U.S. and German curvatures, 1999:01-2018:01.



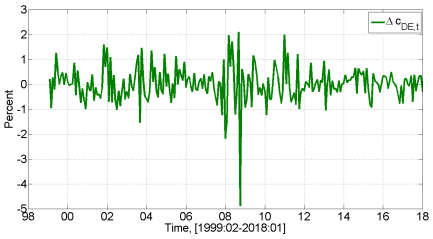
(a) U.S. and German curvatures, in levels.



(b) U.S. and German curvatures, in first differences.



(c) U.S. curvature, in first differences.



(d) German curvature, in first differences.

between the U.S. and German yield curve drivers. In the cross-correlation plots[4] of the U.S. and German country levels in first differences, we observe well-defined peaks at short lags, indicating strong short-run feedback between the two curves. Similar patterns arise for slopes, curvatures, and mixed pairs of estimated country factors. Overall, the evidence points to tight cross-country linkages and non-negligible lead-lag structures in both directions.

We find that the U.S. and German slopes are stationary, positively correlated, and potentially Granger-causing each other. The data-generating process best capturing these properties is the following vector autoregressive (VAR) model of order 5,[5]

$$\begin{bmatrix} s_{US,t} \\ s_{DE,t} \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} + \sum_{i=1}^{5} \mathbf{B}_i \begin{bmatrix} s_{US,t-i} \\ s_{DE,t-i} \end{bmatrix} + \varepsilon_t, \tag{2}$$

where ε_t denotes a vector of reduced-form innovations. Granger causality[6] and impulse response analyses in this VAR(5) model confirm the presence of causality structure between the U.S. and German slopes: each country's slope contains useful information for predicting the other.

Augmented Dickey-Fuller unit root tests for the U.S. and German levels and curvatures fail to reject the unit root hypothesis, suggesting nonstationarity. Johansen tests for cointegration indicate that the U.S. and German levels and curvatures are cointegrated variables. The data-generating process best capturing

[4]These plots are available from the authors on demand.

[5]The lag order of this VAR model was determined in a systematic manner as the lag that minimized Lütkepohl's version of the Akaike Information Criterion (AIC) and for which the residuals of the VAR model were not correlated.

[6]The Granger causality results are available from the authors on demand.

these properties is the following vector error correction (VEC) model of order 3,[7]

$$\begin{bmatrix} \Delta l_{US,t} \\ \Delta l_{DE,t} \\ \Delta c_{US,t} \\ \Delta c_{DE,t} \end{bmatrix} = \boldsymbol{\alpha} \left(\boldsymbol{\beta}' \begin{bmatrix} l_{US,t-1} \\ l_{DE,t-1} \\ c_{US,t-1} \\ c_{DE,t-1} \end{bmatrix} + \mathbf{c}_0 \right) + \sum_{i=1}^{3} \mathbf{B}_i \begin{bmatrix} \Delta l_{US,t-i} \\ \Delta l_{DE,t-i} \\ \Delta c_{US,t-i} \\ \Delta c_{DE,t-i} \end{bmatrix} + \boldsymbol{\varepsilon}_t, \quad (3)$$

where $\boldsymbol{\beta}$ collects the cointegration vectors and $\boldsymbol{\alpha}$ the corresponding adjustment coefficients. Toda-Yamamoto (Toda & Yamamoto (1995)) tests for Granger causality[8] and impulse response analysis suggest the presence of non-trivial causality structure in this VEC(3) model for levels and curvatures, with shocks to U.S. and German factors propagating across borders and feeding back through the error-correction mechanism.

2.3. Outliers and Structural Breaks

Going equation-wise through the VAR(5) model for the slopes and the VEC(3) model for the levels and curvatures, we test for the presence of structural breaks using the methods of Bai & Perron (1998) and Perron et al. (2008).

We find structural breaks in all U.S. and German drivers over the identification sample, with several breaks clustering around the 2007-2008 Financial Crisis and subsequent unconventional monetary policy episodes. Significant changes in the monetary policy of the ECB and the U.S. Fed, which were not anticipated by market participants well in advance, provide a natural explanation for many of these breaks.

[7] The lag of this VEC model was determined using its equivalent VAR representation and following the same procedure explained for the process in (2).

[8] The Granger causality results are available from the authors on demand.

However, the presence of multiple structural breaks in univariate representations of the drivers may also reflect omitted-variable problems: breaks in the univariate dynamics of the U.S. and German levels might be due to missing slopes and curvatures, and vice versa. To obtain a more coherent picture, we therefore test for structural breaks in a multivariate setting and move to a state-space framework that jointly models all U.S. and German yield curve drivers. We develop a "Full"[9] State-Space Model (FSSM) that preserves the VAR dynamics for the slopes and the VEC dynamics for the levels and curvatures within a unified global system. This full model serves as the starting point for our flexible structured state-space models (FSSM, MShock-FSSM, FSSM-US, FSSM-DE) and provides the basis for the outlier and structural-break analysis that underlies the extended-sample forecasting and density evaluation.

3. Modeling Framework

In this section, we develop the FSSM for the US and German yield curve drivers, as well as its variant that accounts for structural alteration, i.e., the MShock-FSSM. State-space models rely on the idea that the object of analysis is a set of r *state variables* which change over time. In the following, we recall mainly from Mittnik (1989); Mittnik (1990); Harvey (1990); Harvey (1993); Hamilton (1994); Commandeur & Koopman (2007); Commandeur et al. (2011); and Durbin & Koopman (2012) the main concepts of state-space modeling.

[9] "Full" in the sense that the model includes all U.S. and German yield curve drivers.

3.1. State-Space Modeling and the Kalman Filter

Let $\boldsymbol{y}_t$ denote an $(n \times 1)$ vector of variables which are actually observed at date t and $\boldsymbol{\xi}_t$ the $(r \times 1)$ vector of *state variables*. The observed variables are related to the *state variables* by a *measurement equation*. The movements of the state variables (equivalently, the *state vector*) are governed by a well-defined process, called the *transition equation*. The measurement equation (5) and the transition equation (4) compose the *state-space representation*[10] of the dynamics of $\boldsymbol{y}$:

$$\underbrace{\boldsymbol{\xi_t}}_{r\times 1} = \underbrace{\mathbf{F}}_{r\times r}\underbrace{\boldsymbol{\xi_{t-1}}}_{r\times 1} + \underbrace{\mathbf{B}}_{r\times r}\underbrace{\mathbf{v_t}}_{r\times 1}, \tag{4}$$

$$\underbrace{\mathbf{y_t}}_{n\times 1} = \underbrace{\mathbf{A'}}_{n\times k}\underbrace{\mathbf{x_t}}_{k\times 1} + \underbrace{\mathbf{H'}}_{n\times r}\underbrace{\boldsymbol{\xi_t}}_{r\times 1} + \underbrace{\mathbf{D}}_{n\times n}\underbrace{\mathbf{w_t}}_{n\times 1} \tag{5}$$

where $\boldsymbol{F}$, $\boldsymbol{B}$, $\boldsymbol{A}'$, and $\boldsymbol{H}'$, and $\boldsymbol{D}$ are matrices of parameters of dimension $(r \times r)$, $(r \times r)$, $(n \times k)$, $(n \times r)$, and $(n \times r)$, respectively. Following Durbin & Koopman (2012), matrix **F** is called the *state transition coefficient matrix*, specifying how the r states are expected to transition from period $t-1$ to t, for all $t = 1, \ldots, T$. Matrix **B** is the *state disturbance loading coefficient matrix*, specifying the additive error model for the state transition from period $t-1$ to t, for all $t = 1, \ldots, T$. Matrix **H** is the *measurement sensitivity coefficient matrix*, specifying how the r states are expected to combine at period t to form the n observations. Matrix **D** is the *observation innovation coefficient matrix*, specifying the error model for the observations for period t, for all $t = 1, \ldots, T$.

Following Hamilton, $\boldsymbol{x}_t$ is a $(k \times 1)$ vector of *exogenous* or *predetermined variables*, in the sense that, $\boldsymbol{x}_t$ provides no information about $\boldsymbol{\xi}_{t+s}$ or $\boldsymbol{w}_{t+s}$ for

[10](Hamilton, 1994, p. 372)

$s = 0, 1, 2, \ldots$ beyond that contained in $\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \ldots, \boldsymbol{y}_1$. $\boldsymbol{x}_t$ could include, for example, lagged values of the measurements $\boldsymbol{y}$ or variables that are uncorrelated with $\boldsymbol{\xi}_t$ and $\boldsymbol{w}_t$ for all τ.

The $(r \times 1)$ vector $\boldsymbol{v}_t$ and the $(n \times 1)$ vector $\boldsymbol{w}_t$ represent the disturbances in the transition and measurement equations, respectively, and are assumed to be vector white noise:

$$E(\boldsymbol{v}_t \boldsymbol{v}_\tau') = \begin{cases} \underbrace{\boldsymbol{Q}}_{r \times r} & \text{for} \quad t = \tau \\ \mathbf{0} & \text{otherwise} \end{cases}$$

$$E(\boldsymbol{w}_t \boldsymbol{w}_\tau') = \begin{cases} \underbrace{\boldsymbol{R}}_{n \times n} & \text{for} \quad t = \tau \\ \mathbf{0} & \text{otherwise} \end{cases}$$

where $\boldsymbol{Q}$ and $\boldsymbol{R}$ are $(r \times r)$ and $(n \times n)$ matrices, respectively. If we also assume that $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ are *unit-variance* white noise processes, their covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ are identity matrices. The disturbances $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ are assumed to be uncorrelated at all lags:

$$E(\boldsymbol{v}_t \boldsymbol{w}_\tau') = 0 \quad \text{for all} \quad t \quad \text{and} \quad \tau.$$

For given values of all system matrices and initial conditions for the state means and covariance matrix, the Kalman filter (Kalman (1960)) enables the estimation of the state vector in three different ways, to produce the *filtered*, the *predicted*, and the *smoothed* estimates of the state vector (Durbin & Koopman (2012); Commandeur et al. (2011)).

3.2. Outliers and Structural Breaks

The adequacy of a state-space model can be assessed by investigating the so-called *auxiliary residuals*, i.e., the *standardized smoothed observation disturbances* (SSODs) and the *standardized smoothed state disturbances* (SSSDs) (Harvey & Koopman (1992); De Jong & Penzer (1998); Commandeur & Koopman (2007); and Commandeur et al. (2011)).

Outliers and structural breaks in state-space models can be dealt with the inclusion of a vector of shocks, $\boldsymbol{\delta}$, to represent the suspected inadequacy in the null state-space model. Theoretically, following De Jong & Penzer (1998), let us assume that an appropriate representation of the process that generates the data $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_n')'$ is the state-space model in 4 and 5. The alternative model is the null model extended to include the vector of shocks, $\boldsymbol{\delta}$. More specifically,

$$\underbrace{\boldsymbol{\xi_t}}_{r\times 1} = \boldsymbol{\Gamma}_t\boldsymbol{\delta} + \underbrace{\mathbf{F}}_{r\times r}\underbrace{\boldsymbol{\xi_{t-1}}}_{r\times 1} + \underbrace{\mathbf{B}}_{r\times r}\underbrace{\mathbf{v_t}}_{r\times 1}, \tag{6}$$

$$\underbrace{\mathbf{y_t}}_{n\times 1} = \boldsymbol{\Lambda}_t\boldsymbol{\delta} + \underbrace{\mathbf{A}'}_{n\times k}\underbrace{\mathbf{x_t}}_{k\times 1} + \underbrace{\mathbf{H}'}_{n\times r}\underbrace{\boldsymbol{\xi_t}}_{r\times 1} + \underbrace{\mathbf{D}}_{n\times n}\underbrace{\mathbf{w_t}}_{n\times 1}, \tag{7}$$

where $\boldsymbol{\Gamma}_t$ and $\boldsymbol{\Lambda}_t$ are called the *shock design matrices* and $\boldsymbol{\delta}$ is the shock magnitude. Outliers, level shifts, and switches can be accounted for by including in the state-space model these intervention variables (also called *intervention signature*).

In regression analysis, a highly accepted idea is that unusual observations may occur in patches. Penzer (2007) explains that, in time series data, a patch of outliers may rise from a level shift, a seasonal break, or any other permanent alteration in structure. To this regard, Penzer argues that patches of unusual behavior should be represented by allowing shocks in the measurement equation of a state-space model.

3.3. FSSM Model Development

The first step in the development of the model consists in casting into state-space form the VAR model for the slopes (Eq. 2) and the VEC model for the levels and curvatures (Eq. 3) to derive the so-called *state-space VAR* (SSVAR) and *state-space VEC* (SSVEC) models. In the next step, we assemble together the SSVAR and the SSVEC models to derive the FSSM. As such, the SSVAR and the SSVEC models represent the sub-models of the FSSM. Following Hamilton (1994), the state equation of the FSSM is:

$$
\underbrace{\begin{bmatrix} \boldsymbol{\xi}_t \\ \boldsymbol{\xi}_{t-1} \\ \boldsymbol{\xi}_{t-2} \\ \boldsymbol{\xi}_{t-3} \\ \boldsymbol{\xi}_{t-4} \\ \boldsymbol{\varphi}_t \\ \boldsymbol{\varphi}_{t-1} \\ \boldsymbol{\varphi}_{t-2} \end{bmatrix}}_{22\times 1} = \underbrace{\left[\begin{array}{cccc|ccc} \mathbf{A}1_{2\times 2} & \mathbf{A}2_{2\times 2} & \dots & \mathbf{A}5_{2\times 2} & & & \\ \mathbf{I}_{2\times 2} & \mathbf{0}_{2\times 2} & \dots & \mathbf{0}_{2\times 2} & & & \\ \mathbf{0}_{2\times 2} & \mathbf{I}_{2\times 2} & \dots & \mathbf{0}_{2\times 2} & & \mathbf{0} & \\ \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times 2} & \dots & \mathbf{0}_{2\times 2} & & & \\ \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times 2} & \dots & \mathbf{0}_{2\times 2} & & & \\ \hline & & & & \mathbf{B}1_{4\times 4} & \mathbf{B}2_{4\times 4} & \mathbf{B}3_{4\times 4} \\ & \mathbf{0} & & & \mathbf{I}_{4\times 4} & \mathbf{0}_{4\times 4} & \mathbf{0}_{4\times 4} \\ & & & & \mathbf{0}_{4\times 4} & \mathbf{I}_{4\times 4} & \mathbf{0}_{4\times 4} \end{array}\right]}_{22\times 22} \underbrace{\begin{bmatrix} \boldsymbol{\xi}_{t-1} \\ \boldsymbol{\xi}_{t-2} \\ \boldsymbol{\xi}_{t-3} \\ \boldsymbol{\xi}_{t-4} \\ \boldsymbol{\xi}_{t-5} \\ \boldsymbol{\varphi}_{t-1} \\ \boldsymbol{\varphi}_{t-2} \\ \boldsymbol{\varphi}_{t-3} \end{bmatrix}}_{22\times 1}
$$

$$
+ \underbrace{\mathbf{B}}_{22\times 22} \underbrace{\begin{bmatrix} \boldsymbol{\epsilon}_t \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ -- \\ \boldsymbol{\varepsilon}_t \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}}_{22\times 1}, \tag{8}
$$

and the measurement equation is:

$$\underbrace{\begin{bmatrix} \boldsymbol{x}_t \\ \Delta \boldsymbol{y}_t \end{bmatrix}}_{6\times 1} = \underbrace{\left[\begin{array}{cccc|ccc} \mathbf{I}_{2\times 2} & \mathbf{0}_{2\times 2} & \dots & \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times 4} & \mathbf{0}_{2\times 4} & \mathbf{0}_{2\times 4} \\ \mathbf{0}_{4\times 2} & \mathbf{0}_{4\times 2} & \dots & \mathbf{0}_{4\times 2} & \boldsymbol{I}_{4\times 4} & \mathbf{0}_{4\times 4} & \mathbf{0}_{4\times 4} \end{array}\right]}_{6\times 22} \underbrace{\begin{bmatrix} \boldsymbol{\xi}_t \\ \boldsymbol{\xi}_{t-1} \\ \boldsymbol{\xi}_{t-2} \\ \boldsymbol{\xi}_{t-3} \\ \boldsymbol{\xi}_{t-4} \\ \boldsymbol{\varphi}_t \\ \boldsymbol{\varphi}_{t-1} \\ \boldsymbol{\varphi}_{t-2} \end{bmatrix}}_{22\times 1}$$

$$+ \underbrace{\begin{bmatrix} \mathbf{0}_{2\times 4} \\ \boldsymbol{\Pi}_{4\times 4} \end{bmatrix}}_{6\times 4} \underbrace{\begin{bmatrix} \boldsymbol{y}_{t-1} \end{bmatrix}}_{4\times 1} + \underbrace{\mathbf{D}}_{6\times 6} \underbrace{\begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{\eta}_t \end{bmatrix}}_{6\times 1}. \tag{9}$$

In Eq. 8, $\boldsymbol{\xi}_t$ denotes a linear combination of current and lagged values of $s_{US,t}$ and $s_{DE,t}$ and $\boldsymbol{\varphi}_t$ denotes a linear combination of current and lagged values of $l_{US,t}, l_{DE,t}, c_{US,t}$, and $c_{DE,t}$, in first differences. Therefore, the state vector of the FSSM holds 22 states, which are the US and German slopes in levels plus 4 lagged states for each of the two slopes and the US and German levels and curvatures in first differences, plus 2 lagged states for each of the two levels and curvatures. The 22-dimensional state equation is collapsed to a 6-dimensional measurement equation, in which $\boldsymbol{x}_t$ denotes the $s_{US,t}$ and $s_{DE,t}$ measurements and $\Delta \boldsymbol{y}_t$ denotes the $l_{US,t}, l_{DE,t}, c_{US,t}$, and $c_{DE,t}$ measurements, in first differences.

Instead of modeling the constant in the VAR model in Eq. 2 as a separate state, we choose to centralize the slopes and, therefore, work with demeaned[11] data in

[11]Working with demeaned data turned out to be a requirement in order to fix issues with

8.

Along the lines of Hamilton (1994); Ribarits & Hanzon (2014a); and Ribarits & Hanzon (2014b), we choose to treat the error-correction term $\mathbf{\Pi} \boldsymbol{y}_{t-1}$ from the VEC model for the levels and curvatures as a *vector of exogenous variables* or a *regression component* in the measurement equation of the SSVEC model.

3.4. Outliers and Structural Breaks in FSSM: The MShock-FSSM

We estimate the FSSM with the Kalman filter and maximum likelihood. A backward pass through the data with the output of the Kalman filter and state and disturbance smoothing algorithms enables the calculation of the standardized smoothed observation disturbances, e_t^* and of the standardized smoothed state disturbances r_t^*, which are analyzed for the presence of outliers and structural breaks. These residuals reveal patches of outliers in almost all yield curve factors predominantly in 2008:08-2008:09, suggesting a sort of synchronicity of outliers across the yield curve drivers.

In the spirit of Penzer (2007), we account for the presence of such patches of outliers in the FSSM with the inclusion of shocks in the measurement equation. We name this enhanced version of FSSM the *MShock-FSSM* model (where M stands for "measurement"). The state equation of MShock-FSSM is equal to that of the FSSM in 8. The intervention variables are incorporated in the regression component of the measurement equation in such a way as to shock separately each component of the measurement vector. More specifically, the *measurement*

numerical optimization failing to converge when estimating the FSSM with the Kalman filter and maximum likelihood.

equation of the MShock-FSSM is defined as:

$$\underbrace{\begin{bmatrix} \boldsymbol{x}_t \\ \Delta \boldsymbol{y}_t \end{bmatrix}}_{6\times 1} = \underbrace{\left[\begin{array}{cccc|ccc} \mathbf{I}_{2\times 2} & \mathbf{0}_{2\times 2} & \dots & \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times 4} & \mathbf{0}_{2\times 4} & \mathbf{0}_{2\times 4} \\ \mathbf{0}_{4\times 2} & \mathbf{0}_{4\times 2} & \dots & \mathbf{0}_{4\times 2} & \boldsymbol{I}_{4\times 4} & \mathbf{0}_{4\times 4} & \mathbf{0}_{4\times 4} \end{array}\right]}_{6\times 22} \underbrace{\begin{bmatrix} \boldsymbol{\xi}_t \\ \boldsymbol{\xi}_{t-1} \\ \boldsymbol{\xi}_{t-2} \\ \boldsymbol{\xi}_{t-3} \\ \boldsymbol{\xi}_{t-4} \\ \boldsymbol{\varphi}_t \\ \boldsymbol{\varphi}_{t-1} \\ \boldsymbol{\varphi}_{t-2} \end{bmatrix}}_{22\times 1} \tag{10}$$

$$+ \underbrace{\begin{bmatrix} \boldsymbol{\delta}_{2\times 2}^{\text{VAR}} & \mathbf{0}_{2\times 4} & \mathbf{0}_{2\times 4} \\ \mathbf{0}_{4\times 2} & \boldsymbol{\delta}_{4\times 4}^{\text{VEC}} & \boldsymbol{\Pi}_{4\times 4} \end{bmatrix}}_{6\times 10} \underbrace{\begin{bmatrix} \boldsymbol{\Lambda}_t^{\text{VAR}} \\ \boldsymbol{\Lambda}_t^{\text{VEC}} \\ \boldsymbol{y}_{t-1} \end{bmatrix}}_{10\times 1} + \underbrace{\mathbf{D}}_{6\times 6} \underbrace{\begin{bmatrix} \boldsymbol{u}_t \\ \boldsymbol{\eta}_t \end{bmatrix}}_{6\times 1},$$

where $\boldsymbol{\Lambda}_t^{\text{VAR}}$ and $\boldsymbol{\Lambda}_t^{\text{VEC}}$ are the shock variables.

The shock variables are equal to 1 or -1 at date points corresponding to outlying measurements, and 0 at all other date points. In order not to undermine the fit of the model, we account only for the most blatant outliers.

4. Point Forecasting

This section evaluates the point forecasting performance of the FSSM and its variants over two distinct exercises: (i) an *identification-sample* analysis based on recursive forecasts within the original sample used for estimation, and (ii) an *extended-sample* evaluation that expands the forecasting period beyond the

identification window. Both exercises are based on multi-horizon recursive forecasts of the U.S. and German Nelson-Siegel level, slope, and curvature factors, subsequently mapped into yields via Eq. (1).

FSSM performance is compared against three widely used benchmarks: (1) the Domestic Diebold-Li model (DDL, Diebold & Li (2006)), (2) the Global Diebold-Li model (GDL, Diebold et al. (2008)), and (3) the Random Walk (RW).

All models are re-estimated every 12 months using the Kalman filter and maximum likelihood, while forecasts are produced at the 1-, 2-, 3-, 6-, 12-, 24-, and 36-month horizons. Forecast accuracy is first assessed using the root mean squared error (RMSE). For each model-benchmark pair, we compute the percent difference

$$\Delta\mathrm{RMSE}\% = 100 \times \frac{\mathrm{RMSE}_A - \mathrm{RMSE}_B}{\mathrm{RMSE}_B},$$

which we summarise by the median across maturities for each horizon.

Beyond RMSE comparison, we formally test whether differences in forecast accuracy are statistically significant using the Diebold-Mariano (DM) statistic (Diebold & Mariano (1995)). The results are reported in Tables 3, 4, 3, and 4. To provide an interpretable and compact summary, the tables report:

- the median $\Delta\mathrm{RMSE}\%$ of our model relative to each benchmark,
- the number of cases (maturities/forecast horizons) at which the DM test rejects the null in favour of the FSSM and its variants at the 5% and 0.1% significance levels. We denote these as "DM Wins $(p < 0.05 \,/\, p < 0.001)$".

To assess the predictive relevance of the curvature factor, we also evaluate two reduced-form specifications: FSSM^{LS} and MShock-FSSM^{LS}, which exclude

the curvature block and model only U.S. and German levels and slopes.[12] Furthermore, we verify whether the inclusion of international information improves the forecasting performance of the FSSM. We do so by forecasting the German term structure with the FSSM-DE model, which is the FSSM model with zero restrictions on the US parameters.

4.1. Identification-Sample Results

Across the identification sample, several patterns emerge.

Short horizons (1-3 months). At short horizons, no model is able to beat the Random Walk, which remains the dominant benchmark in line with the well-known difficulty of forecasting high-frequency yield changes. The FSSM does not outperform the DDL at these horizons and the LS-variants (which exclude curvature) perform markedly worse. However, both FSSM and Mshock-FSSM consistently outperform the GDL. This behaviour may be explained by the fact that short-run dynamics are governed by high-frequency noise, announcement effects, and short-lived policy surprises. During these horizons, yield curves display limited exploitable structure; as a result, models relying on global or long-run dynamics—such as FSSM or GDL—cannot overcome the RW benchmark. The LS deterioration confirms that curvature contains valuable short-run information even within the identification period.

Medium horizons (6-12 months). At medium horizons, the gains from modeling structure become visible. Both FSSM and Mshock-FSSM show clear and systematic improvements over the GDL benchmark and begin to narrow the gap relative

[12] The mathematical definition of these models is available from the authors on demand.

to DDL. LS-variants remain weak, particularly for Germany, confirming that curvature contributes meaningfully to predictive content during this period. Improvements vs. GDL highlight the relevance of cross-country dynamic linkages: the FSSM captures spillovers in slopes and long-run factor interactions that GDL's simpler factor structure misses. In Germany, the moderate but positive improvements reflect the fragmented euro-area monetary environment of the pre-2018 period (asymmetric ECB stance, sovereign risk episodes, and structural idiosyncrasies), which limited the power of global drivers at medium horizons.

Long horizons (2-3 years). There are robust improvements for U.S. vs GDL and DDL, stemming from the models' ability to incorporate persistent global co-movements. For Germany, there are minor gains vs DDL and the GDL remains competitive. The RW remains hard to beat. LS-variants remain reasonably competitive but still underperform the full model, particularly in Germany. These observed behaviors reflect the FSSM's structural design: its VEC dynamics impose long-run cointegration between U.S. and German level and curvature factors, enabling persistent global co-movements to be translated into more accurate long-horizon forecasts. The GDL lacks the full error-correction structure and underestimates slow-moving global equilibrium forces. For Germany, the outperformance of the full FSSM (relative to FSSM-DE) confirms the importance of U.S. monetary spillovers—the Fed acted as the leading source of global yield-curve movements during the identification sample, with the ECB reacting more gradually.

4.2. Extended-Sample Results

The extended-sample evaluation—covering a much longer period and including episodes of policy-rate normalization, the pandemic, and the subsequent

Table 1: Point forecasting results: Identification sample, U.S..

Horizon	Model	Med ΔRMSE vs RW (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs RW	Med ΔRMSE vs DDL (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs DDL	Med ΔRMSE vs GDL (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs GDL
Short	**FSSM**	22.65%	0 / 0	19.89%	0 / 0	**-46.94%**	**21 / 17**
Short	**Mshock-FSSM**	22.02%	0 / 0	18.17%	0 / 0	**-46.07%**	**21 / 17**
Short	Mshock-FSSM-LS	203.56%	0 / 0	107.02%	0 / 0	27.02%	**2 / 1**
Short	FSSM-LS	209.06%	0 / 0	183.27%	0 / 0	30.23%	**2 / 1**
Medium	**Mshock-FSSM**	28.85%	0 / 0	3.41%	**2 / 0**	**-61.02%**	**21 / 20**
Medium	**FSSM**	32.79%	0 / 0	4.64%	**1 / 0**	**-60.15%**	**21 / 20**
Medium	Mshock-FSSM-LS	57.98%	0 / 0	24.49%	0 / 0	**-45.54%**	**21 / 19**
Medium	FSSM-LS	60.64%	0 / 0	28.42%	0 / 0	**-44.40%**	**21 / 19**
Long	**Mshock-FSSM**	28.68%	0 / 0	**-9.43%**	**10 / 1**	**-60.18%**	**14 / 14**
Long	**FSSM**	31.74%	0 / 0	**-32.09%**	**12 / 0**	**-59.15%**	**14 / 14**
Long	Mshock-FSSM-LS	53.98%	0 / 0	**-22.75%**	**10 / 0**	**-51.36%**	**14 / 14**
Long	FSSM-LS	55.75%	0 / 0	**-21.44%**	**8 / 0**	**-50.80%**	**14 / 14**

Table 2: Point forecasting results: Identification sample, Germany.

Horizon	Model	Med ΔRMSE vs RW (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs RW	Med ΔRMSE vs DDL (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs DDL	Med ΔRMSE vs GDL (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs GDL
Short	**Mshock-FSSM**	14.17%	0 / 0	1.82%	0 / 0	**-25.36%**	**20 / 13**
Short	**FSSM**	14.24%	0 / 0	5.71%	0 / 0	**-24.62%**	**20 / 13**
Short	Mshock-FSSM-LS	168.62%	0 / 0	140.65%	0 / 0	75.13%	**3 / 3**
Short	FSSM-LS	171.85%	0 / 0	143.46%	0 / 0	76.95%	**3 / 3**
Medium	**Mshock-FSSM**	18.68%	0 / 0	1.93%	**1 / 0**	**-11.76%**	**11 / 7**
Medium	**FSSM**	26.41%	0 / 0	5.26%	0 / 0	**-5.66%**	**9 / 6**
Medium	Mshock-FSSM-LS	67.87%	0 / 0	34.21%	0 / 0	16.41%	**3 / 3**
Medium	FSSM-LS	71.47%	0 / 0	38.01%	0 / 0	21.14%	**3 / 3**
Long	**Mshock-FSSM**	21.09%	0 / 0	**-14.66%**	**11 / 7**	11.46%	**4 / 2**
Long	**FSSM**	24.37%	0 / 0	**-12.40%**	**9 / 6**	13.65%	**3 / 2**
Long	Mshock-FSSM-LS	16.49%	0 / 0	**-17.90%**	**10 / 10**	10.93%	**2 / 2**
Long	FSSM-LS	17.04%	0 / 0	**-17.95%**	**10 / 9**	11.35%	**2 / 2**

tightening cycle—delivers results that are fully consistent with, but sharper than, the identification-sample findings:

Short horizons (1-3 months). Compared to the identification sample, the performance of our models in the U.S. case remains strong vs GDL, but slightly weaker due to higher noise and regime changes. In the case of Germany, however, the stronger integration with the global factors improves the performance vs GDL and DDL.

Medium horizons (6-12 months). All our models show clear, significant gains vs GDL for Germany. This is supported by more DM wins and it is in line with the ECB-Fed synchronization that enhances the predictability of the German term structure. In the case of U.S., the performance of FSSM and its variants is still very strong vs GDL, though less stable and there are still no improvements vs the RW or DDL. The short- and medium-term linkages to the global factors remain predictive but slightly weakened due to more idiosynchratic U.S. dynamics.

Long horizons (2-3 years). The performance for the U.S. deteriorates sharply, with fewer DM wins and smaller gains vs GDL. These results reflect that the increased macro volatility and structural shifts post-2018 (trade tensions, COVID-19 disruptions, rapid rate cuts followed by sharp hikes, and strong fiscal shocks) weakened the long-term predictability of the U.S. term structure. In other words, at longer horizons, structural breaks and regime shifts in the U.S. eroded the ability of historical cointegration and persistence structures to forecast yields effectively. On the contrary, in the case of Germany, we observe that the LS and full models yield significant gains vs GDL, showing stronger long-term linkages. Post-2018, the German yield curve became more globally synchronized due to ECB

policy normalization followed by aggressive easing during the pandemic, deeply negative rates giving way to rapid hikes during 2022-2023 to combat inflation, rising global risk correlations, and higher sensitivity to US yield dynamics. The extended sample once again shows that FSSM outperforms its Germany-only variant FSSM-DE. The empirical interpretation is clear: U.S. monetary policy shifts lead euro-area yield movements, and integrating this cross-country dynamic improves the long-horizon accuracy of German forecasts.

Table 3: Point forecasting results: Extended sample, U.S..

Horizon	Model	Med ΔRMSE vs RW (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs RW	Med ΔRMSE vs DDL (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs DDL	Med ΔRMSE vs GDL (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs GDL	Change (Ext vs ID)
Short	**FSSM**	18.65	0/0	13.05	0/0	**-44.39**	**20/17**	**Improved vs RW. Improved vs DDL.** Worse vs GDL.
Short	**Mshock-FSSM**	17.83	0/0	11.76	0/0	**-42.07**	**21/18**	**Improved vs RW. Improved vs DDL.** Worse vs GDL.
Short	Mshock-FSSM-LS	147.55	0/0	127.01	0/0	21.68	2/2	**Improved vs RW.** Worse vs DDL. Stable vs GDL.
Short	FSSM-LS	151.23	0/0	130.39	0/0	22.72	2/2	**Improved vs RW. Improved vs DDL.** Stable vs GDL.
Medium	**Mshock-FSSM**	32.87	0/0	18.96	0/0	**-46.92**	**21/18**	Worse vs RW. Worse vs DDL. Worse vs GDL.
Medium	**FSSM**	33.70	0/0	19.87	0/0	**-46.63**	**21/18**	Stable vs RW. Worse vs DDL. Worse vs GDL.
Medium	**Mshock-FSSM-LS**	56.29	0/0	36.24	0/0	**-34.59**	**21/18**	Stable vs RW. Worse vs DDL. Worse vs GDL.
Medium	**FSSM-LS**	57.86	0/0	38.20	0/0	**-33.65**	**21/18**	**Improved vs RW.** Worse vs DDL. Worse vs GDL.
Long	**Mshock-FSSM**	87.07	0/0	58.20	0/0	**-10.63**	**6/4**	Worse vs RW. Worse vs DDL. **Fewer wins vs GDL.**
Long	**FSSM**	87.99	0/0	28.75	0/0	**-10.20**	**6/4**	Worse vs RW. Worse vs DDL. **Fewer wins vs GDL.**
Long	**Mshock-FSSM-LS**	95.17	0/0	34.63	0/0	**-7.07**	**6/4**	Worse vs RW. Worse vs DDL. **Fewer wins vs GDL.**
Long	**FSSM-LS**	95.77	0/0	34.93	0/0	**-6.74**	**6/4**	Worse vs RW. Worse vs DDL. **Fewer wins vs GDL.**

Table 4: Point forecasting results: Extended sample, Germany.

Horizon	Model	Med ΔRMSE vs RW (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs RW	Med ΔRMSE vs DDL (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs DDL	Med ΔRMSE vs GDL (%)	DM Wins ($p < 0.05$ / $p < 0.001$) vs GDL	Change (Ext vs ID)
Short	**Mshock-FSSM**	17.37	0/0	**-2.42**	**5/2**	**-41.66**	**21/19**	Worse vs RW. **Improved vs DDL. More significant wins vs GDL.**
Short	**FSSM**	19.44	0/0	**-2.64**	**6/1**	**-40.26**	**21/20**	Worse vs RW. **Improved vs DDL. More significant wins vs GDL.**
Short	Mshock-FSSM-LS	166.01	0/0	113.53	0/0	41.48	5/4	**Improved vs RW. Improved vs DDL.** Stable vs GDL.
Short	FSSM-LS	168.76	0/0	115.42	0/0	42.74	5/4	**Improved vs RW. Improved vs DDL.** Stable vs GDL.
Medium	**Mshock-FSSM**	23.65	0/0	**-6.68**	**7/1**	**-34.07**	**21/21**	Worse vs RW. **Improved vs DDL. More significant wins vs GDL.**
Medium	**FSSM**	26.67	0/0	**-4.83**	**8/0**	**-31.77**	**21/21**	Stable vs RW. **Improved vs DDL. More significant wins vs GDL.**
Medium	**Mshock-FSSM-LS**	48.16	0/0	8.15	0/0	**-23.36**	**17/14**	**Improved vs RW. Improved vs DDL. More significant wins vs GDL.**
Medium	**FSSM-LS**	51.24	0/0	9.69	0/0	**-21.59**	**17/13**	**Improved vs RW. Improved vs DDL. More significant wins vs GDL.**
Long	**Mshock-FSSM-LS**	32.31	0/0	**-10.25**	**5/0**	**-21.27**	**14/14**	Worse vs RW. Worse vs DDL. **More significant wins** vs GDL.
Long	**FSSM-LS**	32.28	0/0	**-10.01**	**5/0**	**-21.17**	**14/14**	Worse vs RW. Worse vs DDL. **More significant wins** vs GDL.
Long	**Mshock-FSSM**	35.50	0/0	**-7.61**	**3/0**	**-18.63**	**12/11**	Worse vs RW. Worse vs DDL. **More significant wins** vs GDL.
Long	**FSSM**	41.99	0/0	**-3.19**	**4/0**	**-15.68**	**10/8**	Worse vs RW. Worse vs DDL. **More significant wins** vs GDL.

4.3. Summary

Taken together, the identification-sample and extended-sample results paint the following coherent picture. At short to medium horizons (1-12 months), the RW remains the benchmark to beat, DDL is competitive, and FSSM is comparable but not superior. At long horizons (2-3 years), FSSM delivers its strongest performance and outperforms both Diebold-Li models. These gains originate from correctly modeling the VEC dynamics and the cross-country error-correction forces. The curvature matters: removing the curvature factor (FSSM^{LS} and MShock-FSSM^{LS}) materially worsens performance at all horizons. Global information matters: FSSM-DE, which removes U.S. information when forecasting Germany, performs consistently worse, confirming that U.S. yield dynamics contain predictive content for German yields.

Overall, the FSSM is not designed to beat the RW at very short horizons—no structural model does—but it offers clear gains over existing factor models at medium and especially long horizons. These long-horizon advantages are economically meaningful for applications such as valuation, risk management, and macro-finance scenario design, where horizon lengths of 1-3 years are most relevant.

5. Directional Accuracy

Point forecast accuracy, typically assessed through RMSE, is only one dimension of model performance. For many financial and policy applications the economically relevant question is not how large the forecast error is, but whether the model correctly predicts the *direction* of change. This is particularly true for trading strategies based on curve steepening/flattening views, tactical duration

and slope tilts, issuance timing, and forward-looking risk management.

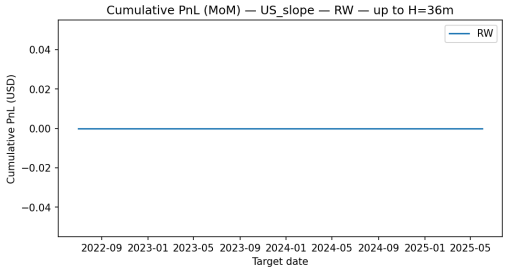
This raises an important question: *If sophisticated models such as the FSSM do not systematically beat the Random Walk (RW) in RMSE, why should one use them at all?* Our results show that the answer lies in directional forecasting performance and its associated economic value: the FSSM delivers substantially better directional signals than both the RW and the GDL model, even in situations where point-forecast RMSE differences are small. More specifically, analyzing the predicted direction of the U.S. and German slopes in a steepening/flattening trading exercise[13], we observe that the RW has very competitive RMSE, reflecting high persistence and the dominance of high-frequency noise. However, its forecasts imply no change in the slopes at all times, so it offers no exploitable directional information. As such, the RW produces no trades, zero cumulative PnL (Figures 5c and 5b), and zero Sharpe.

For both the U.S. and Germany, the FSSM achieves remarkable directional hit rates for the slope factor. The model generates actionable up/down signals and it captures turning points that the RW misses. As seen in Figures 5e and 5f, the cumulative PnL at the last forecast horizon is positive ($\sim \$2.1M$ (DE) and $\sim \$1.9M$ (US) on \$1M notional), as well as the Sharpe ratio.

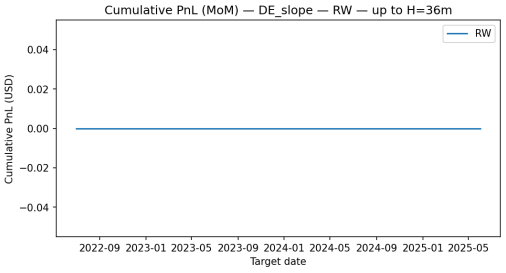
Despite competitive RMSE performance, the GDL model performs systematically worse in directional forecasting of slope changes. Its implied signals frequently take the wrong sign at turning points, which results in negative cumulative PnL ($\sim \$\text{-}2.8M$ (DE) and $\sim \$\text{-}1.3M$ (US) on \$1M notional, Figures 5c and 5f), even when the level forecasts themselves are accurate in an RMSE sense.

[13]Explained in more detail in Appendix Appendix B.

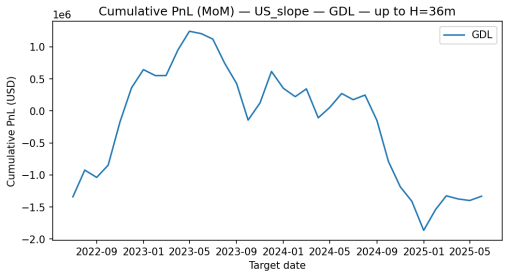
Figure 5: Directional evaluation results: Cumulative PnL for the slope factor.



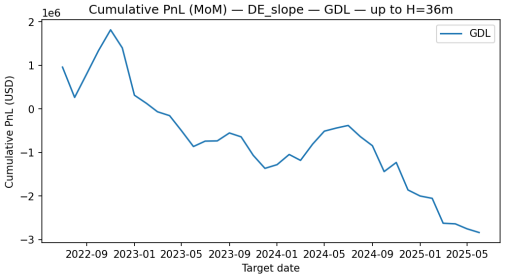
(a) RW: Cum PnL, $s_{US,t}$, h=36M



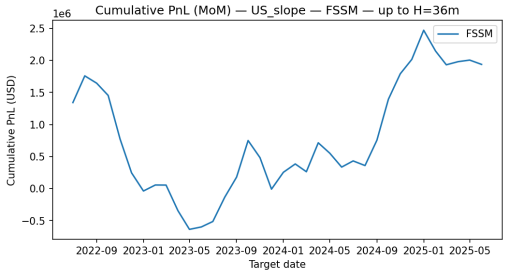
(b) RW: Cum PnL, $s_{DE,t}$, h=36M



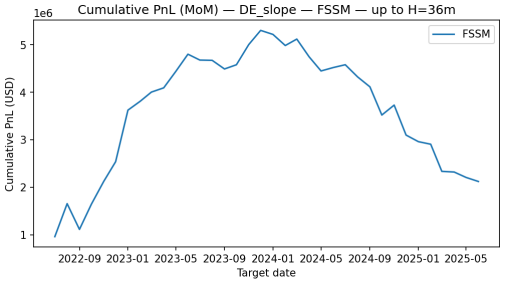
(c) GDL: Cum PnL, $s_{US,t}$, h=36M



(d) GDL: Cum PnL, $s_{DE,t}$, h=36M



(e) FSSM: Cum PnL, $s_{US,t}$, h=36M



(f) FSSM: Cum PnL, $s_{DE,t}$, h=36M

These results demonstrate that RMSE rankings and economic value can diverge sharply. The RW minimizes squared forecast errors but is essentially useless for directional trading. The FSSM, by contrast, yields economically meaningful directional signals and robust positive PnL, even when its RMSE is only comparable to or slightly worse than the RW. Taken together, the evidence supports a clear conclusion. In economically relevant applications, forecasting the *direction* of yield curve movements is more important than minimizing RMSE. Structural models such as the FSSM, which explicitly capture the dynamics of yield curve drivers and their cross-country linkages, provide precisely this directional information, whereas RW and GDL do not. Thus, even though the RW remains a formidable benchmark in traditional RMSE comparisons, the FSSM retains substantial practical value through its superior directional accuracy and the positive trading profits it generates.

6. Density Forecasting: Bootstrapping Method

This section evaluates the full predictive *densities* generated by the competing models. Whereas Section 4 focused on point forecasts, we are now interested in the entire distribution of future yield curves, which is what matters for risk management, fan charts, issuance strategy and stress-testing.

All models are put on a common footing via a residual bootstrap[14] procedure applied to the out-of-sample forecast errors. This produces bootstrapping draws for the whole yield curve at each forecast origin and horizon, which are then calibrated and scored using proper multivariate scoring metrics (i.e., Coverage,

[14]The notation and pipeline of our bootstrap approach is explained in Appendix

Average Predictive Intervals (PIs) Width, RMSE of PIs, Joint Log Scores (JLS), and Energy Score (ES)).

6.1. Results for the U.S.

Figures 6, 7, and 8 summarize[15] the bootstrap performance for the U.S. across all maturities and horizons.

The calibrated interval widths show a clear pattern. RW produces the narrowest bands on average, GDL widens them moderately, and FSSM/FSSM-US deliver the widest intervals, especially at the 2-3 year horizons. Wider intervals are not a defect. They reflect FSSM's ability to incorporate regime shifts (ZLB exit, 2020 pandemic shocks, Fed QT), cross-country cointegration, richer slope/curvature dynamics, and volatility breaks in U.S. term premia. RW and GDL ignore many of these channels, yielding overly tight, under-dispersed predictions. In practice, risk managers using RW/GDL would be overconfident and policy institutions would underestimate the width of likely future rate paths. FSSM fan-charts would better match the actual dispersion observed during tightening cycles.

The RMSE heatmaps for PI centers show that FSSM and FSSM-US remain competitive with RW and GDL in terms of point accuracy implied by the intervals. There is no evidence that the wider bands come at the cost of severely biased centers.

The multivariate scores in Figure 8 are more revealing. For the U.S., at medium and long horizons (24-36 months), the FSSM achieves the highest JLS values. RW's JLS deteriorates markedly as the horizon lengthens, reflecting the fact that a pure

[15]Further heatmaps with Coverage, PI Width, and RMSE of PI Width at 90%, 95%, and 99% are reported in Appendix C.

persistence model fails to capture the evolving co-movements of the curve. GDL performs reasonably but is consistently dominated by FSSM in the 1-3 year range. With regards to the ES, the RW delivers the best scores, followed by the GDL, and the FSSM.

Taken together, these findings show that the FSSM produces the most coherent and realistic joint predictive densities for the U.S. yield curve, especially beyond one year. For applications such as macro-financial fan charts, scenario design, and long-horizon issuance or hedging decisions, these joint scores are more informative than marginal RMSE alone.

The FSSM explicitly imposes cointegration across levels and curvatures and links U.S. and German factors. In a period with large global shocks, QE/QT cycles and shifting term premia, this structure allows the model to capture persistent global co-movements as well as regime changes. The resulting densities are conservative but well calibrated, making them particularly useful for designing probability bands around policy-rate and yield-curve projections, stress-testing bank and insurance portfolios under curve-wide shocks, and evaluating the distribution of future term premia relevant for long-dated issuance. RW and GDL remain attractive for very short-horizon tactical uses, but they underperform FSSM once multi-quarter decisions are considered.

The purely domestic specification FSSM-US, which imposes zero loadings on all German terms, performs markedly worse than the full FSSM. This suggests that, over our sample, U.S. term-structure dynamics are tightly linked to global factors that are also reflected in German yields. In the FSSM those linkages enter the VEC structure and improve the filtering of common level and curvature factors. Imposing zero foreign coefficients breaks the long-run error-correction

mechanism and reduces cross-sectional information in the Kalman filter, leading to noisier state estimates and inferior medium- and long-horizon forecasts. In other words, even if the policy interest is purely domestic, exploiting foreign yield-curve information is empirically valuable for forecasting the U.S. curve.

Figure 6: Boostrapping results for RW and GDL and country U.S.: Average 95% PI width and its RMSE.

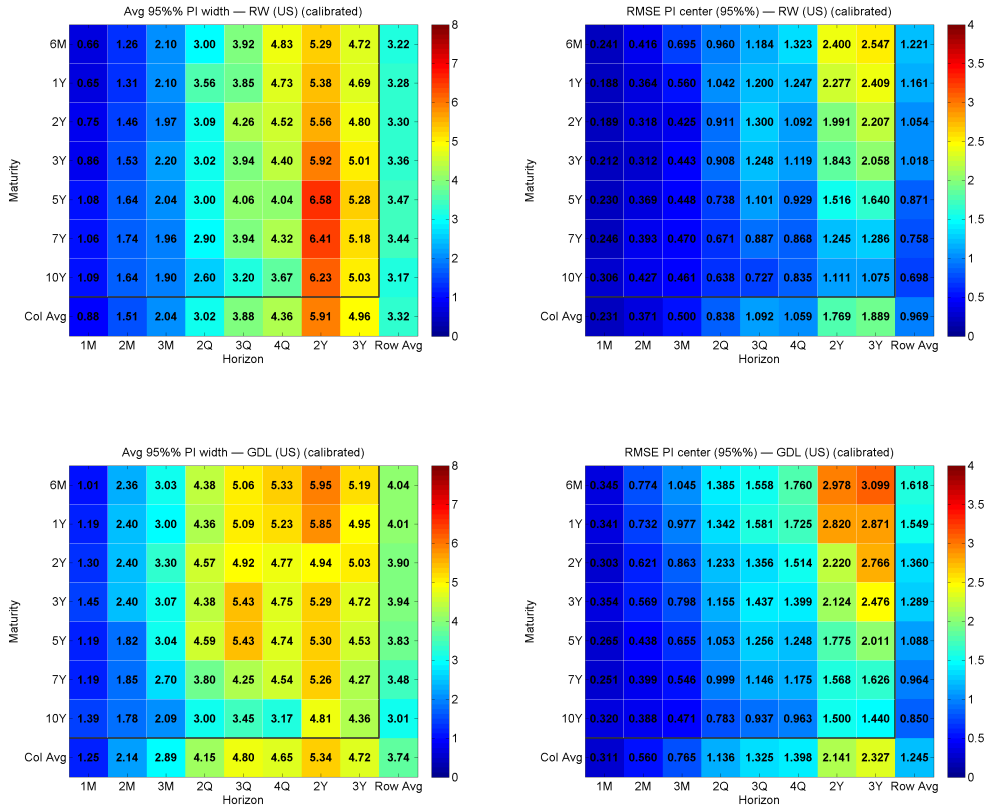


Figure 7: Boostrapping results for FSSM and FSSM-US and country U.S.: Average 95% PI width and its RMSE.

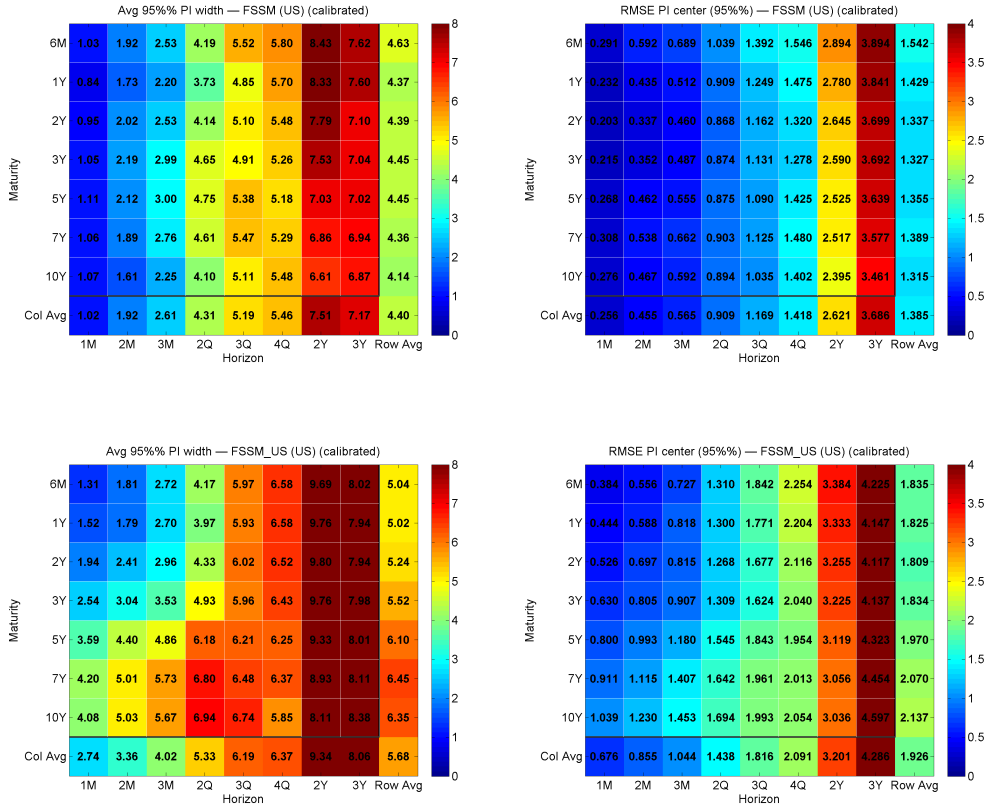
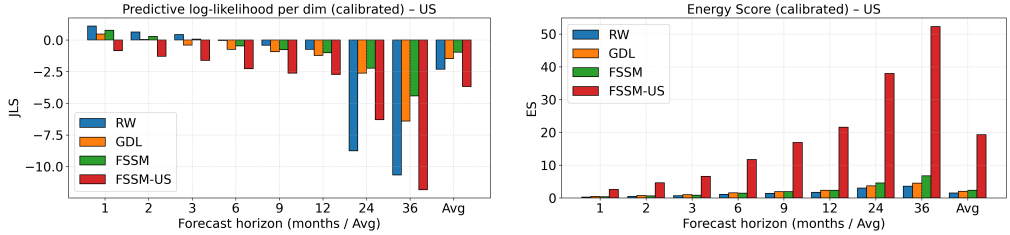


Figure 8: Boostrapping results for country U.S. and all models: Joint Log Scores and Energy Scores.



6.2. Results for Germany

Figures 9, 10, and 11 summarize[16] report the corresponding results for Germany.

For Germany, RW again produces the narrowest calibrated intervals with. GDL bands are slightly wider and remain competitive up to 2-3 years. The FSSM and FSSM-DE variants generate wider intervals, particularly at the longest horizons, reflecting the added uncertainty from modeling explicit U.S.-DE linkages and euro-area regime changes. The RMSE of PI centers indicates that FSSM is well-centered at short and medium horizons, with only modest deterioration at 2-3 years relative to the simpler models.

The JLS and ES patterns suggest a nuanced ranking. The RW and GDL are neck-and-neck overall, while the FSSM and FSSM-DE lag across both metrics. GDL has an edge in JLS, reflecting the flexibility of its global factor structure when euro-area dynamics drift away from tight U.S.-DE cointegration. The FSSM-DE's poor performance vs the FSSM proves once again that the inclusion of U.S. factors does improve the forecastability of the German term structure.

In general, for German yields, the results are consistent with a story where short-horizon dynamics are dominated by euro-area persistence and ECB policy gradualism, while medium-term outcomes reflect both global forces and idiosyncratic European factors. For 1-4 month horizons, RW (and to a lesser extent GDL) provides sufficiently accurate and sharp densities for desk-level risk control and high-frequency hedging of Bund/Bobl/Schatz positions. For quarter-to-year horizons, FSSM is preferred: it offers well-calibrated intervals and the most coher-

[16]Further heatmaps with Coverage, PI Width, and RMSE of PI Width at 90%, 95%, and 99% are reported in Appendix C.

ent cross-maturity dynamics, which is important for DMO[17] tenor and auction planning, bank ALM[18] bands for NII/EVE[19], and ECB communication ranges. For 2-3 year policy and issuance views, GDL's slightly better joint fit at the long end makes it a useful complement, although its lack of explicit cointegration restrictions suggests using it alongside, rather than instead of, the structurally richer FSSM.

6.3. Summary

The residual bootstrap analysis shows that, once all models are calibrated to comparable marginal coverage, differences in density performance are driven by their structural treatment of dynamics and cross-maturity co-movements. For the U.S., the FSSM delivers the sharpest and most coherent joint predictive densities, particularly at policy-relevant horizons beyond one year. For Germany, FSSM excels at medium horizons, while RW and GDL retain an edge at very short and very long horizons, respectively.

For monetary-policy analysis, macro-financial risk management, and yield-curve scenario design, these results argue strongly in favor of structurally rich state-space models such as FSSM, even when simpler benchmarks remain competitive in point-forecast RMSE.

[17] DMO: Debt Management Office.

[18] ALM: Asset-Liability Management.

[19] NII: Net Interest Income. EVE: Economic Value of Equity.

Figure 9: Boostrapping results for RW and GDL and country Germany: Average 95% PI width and its RMSE.

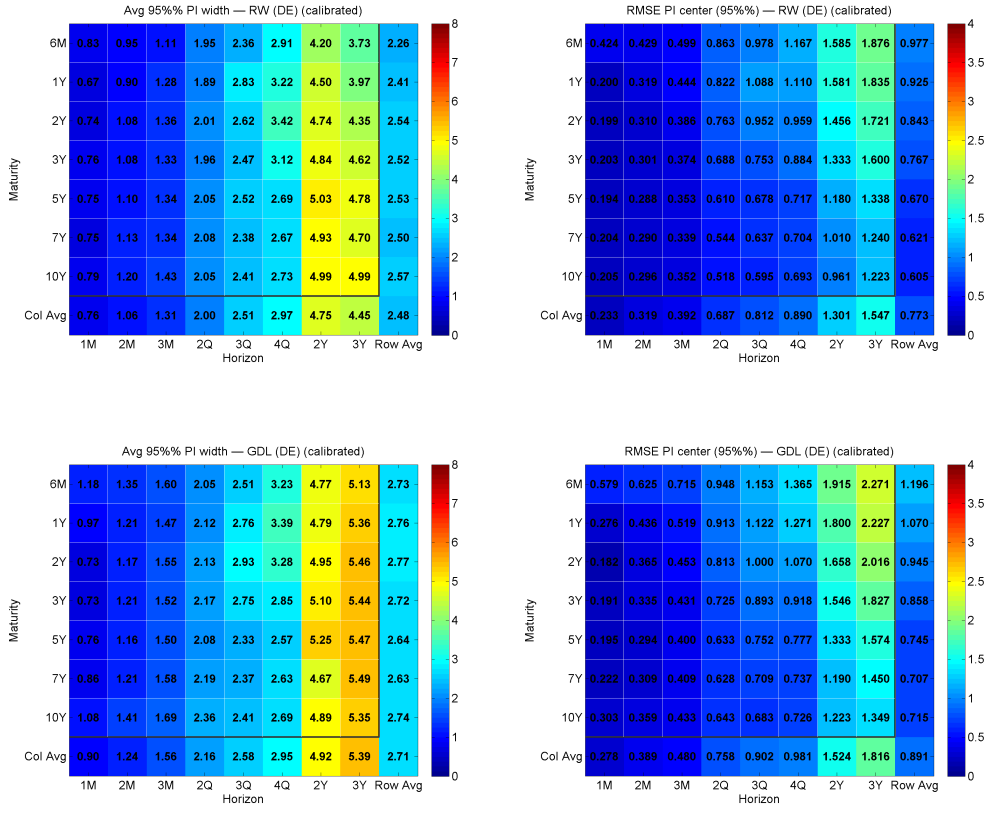


Figure 10: Boostrapping results for FSSM and FSSM-DE and country Germany: Average 95% PI width and its RMSE.

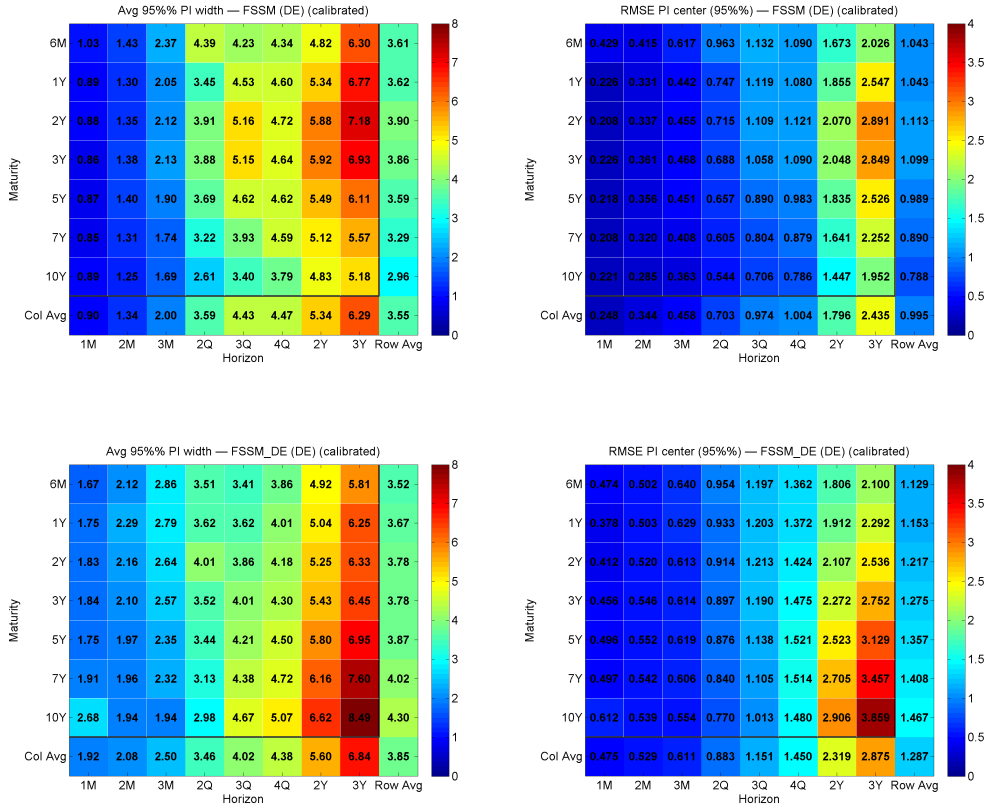
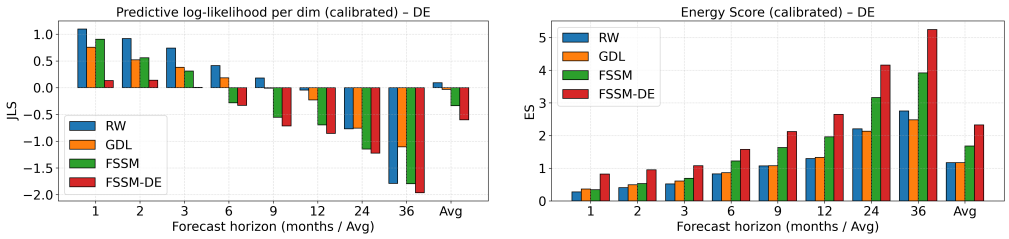


Figure 11: Boostrapping results for country Germany and all models: Joint Log Scores and Energy Scores.



7. Density Forecasting: Monte Carlo Method

This section evaluates density forecasts generated through a fully model-driven Monte Carlo (MC) procedure[20]. Unlike the bootstrap approach—which resamples empirical forecast errors—the MC method draws directly from each modelfis state-space innovations, thereby assessing the predictive distribution implied by the *structural dynamics* of FSSM, GDL, and RW.

7.1. Results for the U.S.

Figures 12, 13, and 14 summarize[21] the MC performance for the U.S. across all maturities and horizons.

The FSSM yields the sharpest PIs and lowest RMSE, followed by the GDL, while the RW produces the widest PIs and highest RMSE. RW's "last-value" dynamics imply no structural propagation. Multi-quarter uncertainty expands mechanically, producing very wide PIs, the lowest JLS, and the worst ES. RW is

[20]Our MC pipeline is explained in greater detail in Appendix Appendix D.

[21]Further heatmaps with Coverage, PI Width, and RMSE of PI Width at 90%, 95%, and 99% are reported in Appendix D.

Figure 12: Monte Carlo results for RW and GDL and country U.S.: Average 95% PI width and its RMSE.

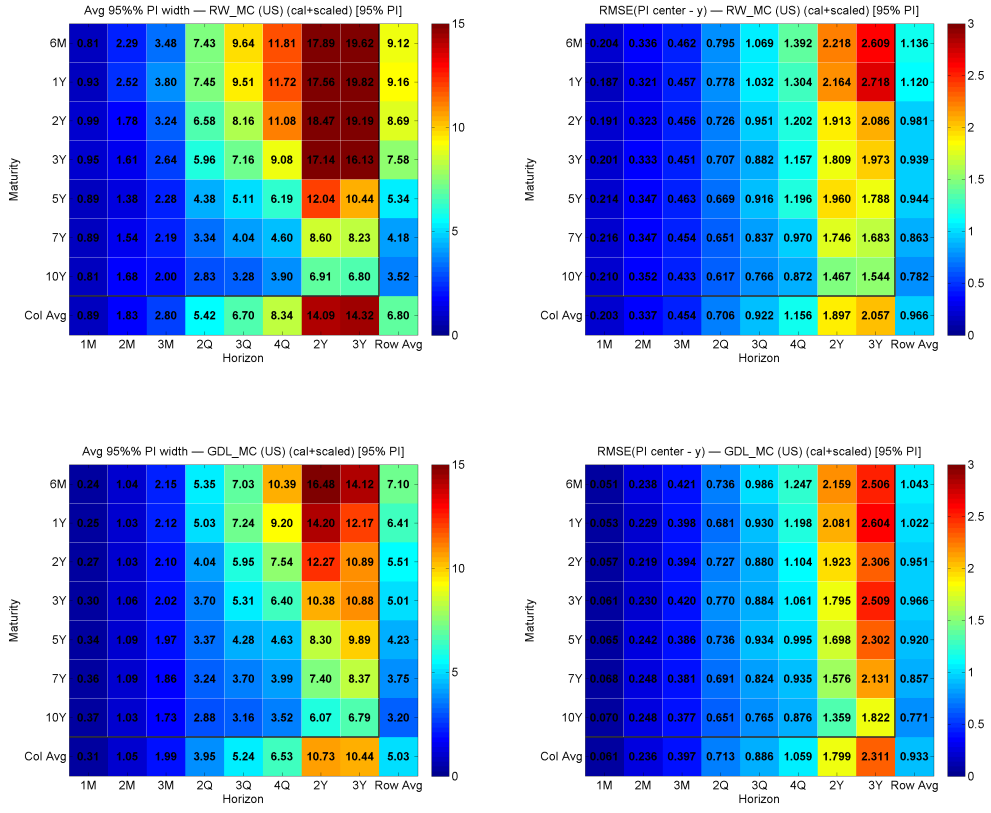


Figure 13: Monte Carlo results for FSSM and FSSM-US and country U.S.: Average 95% PI width and its RMSE.

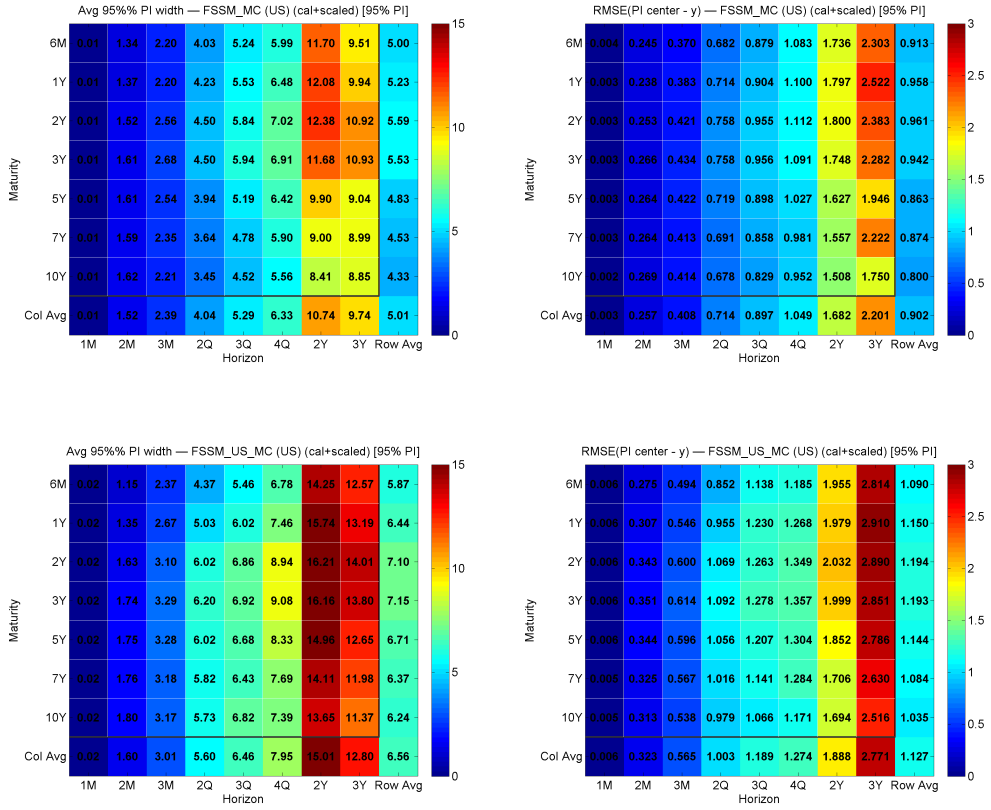
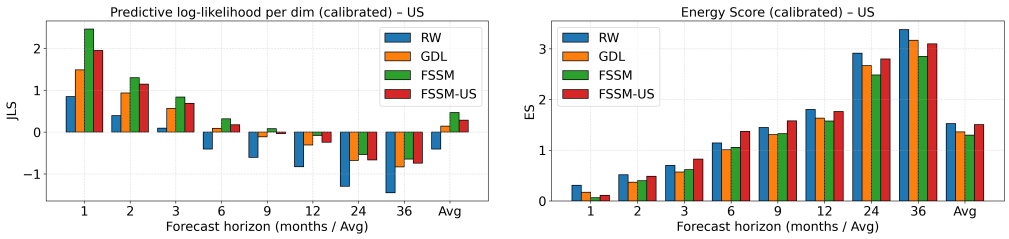


Figure 14: Monte Carlo results for country U.S. and all models: Joint Log Scores and Energy Scores.



adequate only when predictability is negligible. In the case of GDL, the global factor structure enforces smooth shape evolution and produces reasonably sharp densities. However, its weak cross-equation restrictions prevent it from fully capturing the joint co-movement of U.S. maturities, resulting in systematically lower JLS than FSSM. In the FSSM, the explicit cointegration between U.S. and German level-curvature factors, combined with a full VAR structure for slopes, yields state evolutions that propagate shocks accurately across maturities and horizons. This produces the highest JLS at all horizons, the lowest ES, and tight but well-calibrated PIs. These features reflect the FSSM's ability to model monetary-policy cycles, cross-country spillovers, and long-run equilibrium corrections.

For very short-run desk decisions (1M horizon), the RW or GDL are sufficient when predictability is intrinsically low. For strategic risk management and scenario generation (3–12M horizons), term-premium analysis and curve-shape trading signals, and ALM, issuance policy, and balance-sheet risk (3M–2Y horizons), the FSSM is preferred. Its superior JLS produces the most reliable medium-term uncertainty bands and joint density structure. As such, the FSSM is capable of producing more credible multi-maturity risk scenarios and the cross-maturity

coherence improves steepener/flattening strategies and macro factor extraction. The FSSM-US variant also performs notably worse than the full FSSM, indicating that German yield curve information contributes meaningfully to U.S. forecast accuracy—consistent with well-documented global term-structure spillovers.

7.2. Results for Germany

Figures 15, 16, and 17 summarize[22] report the corresponding results for Germany.

Similarly to the U.S. results, the FSSM delivers also for Germany the sharpest PIs and lowest RMSE, the GDL follows, while the RW lags behind. For the RW, the persistence in euro-area rates yields decent 1-3M fit, but without structural linkages its joint density deteriorates beyond one year—lowest ES and lowest JLS. The flexible global factor dynamics in GDL adapt well to medium-term fluctuations, especially during periods of U.S.–euro-area divergence. However, the lack of error-correction structure limits its joint coherence. The cointegrated U.S.–German level–curvature block and the VAR slope structure in the FSSM generate the highest JLS at all horizons, lowest ES, and the tightest and most coherent joint densities. The FSSM captures the empirically important fact that ECB cycles often lag the Federal Reserve, and German yields adjust partly through global spillovers.

Given these results, the FSSM emerges as the most suitable framework for Germany across virtually all forecasting horizons. At the short end (1-3 months), it delivers the strongest probabilistic performance—achieving both the lowest ES and

[22]Further heatmaps with Coverage, PI Width, and RMSE of PI Width at 90%, 95%, and 99% are reported in Appendix C.

Figure 15: Monte Carlo results for RW and GDL and country Germany: Average 95% PI width and its RMSE.

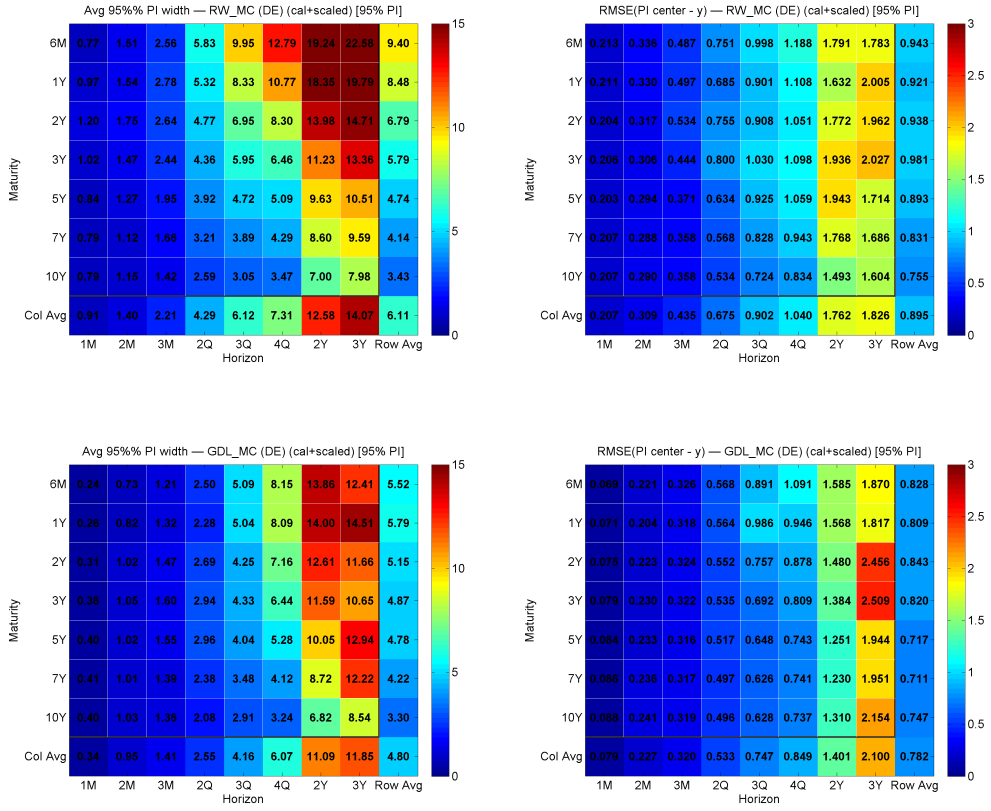


Figure 16: Monte Carlo results for FSSM and FSSM-DE and country Germany: Average 95% PI width and its RMSE.

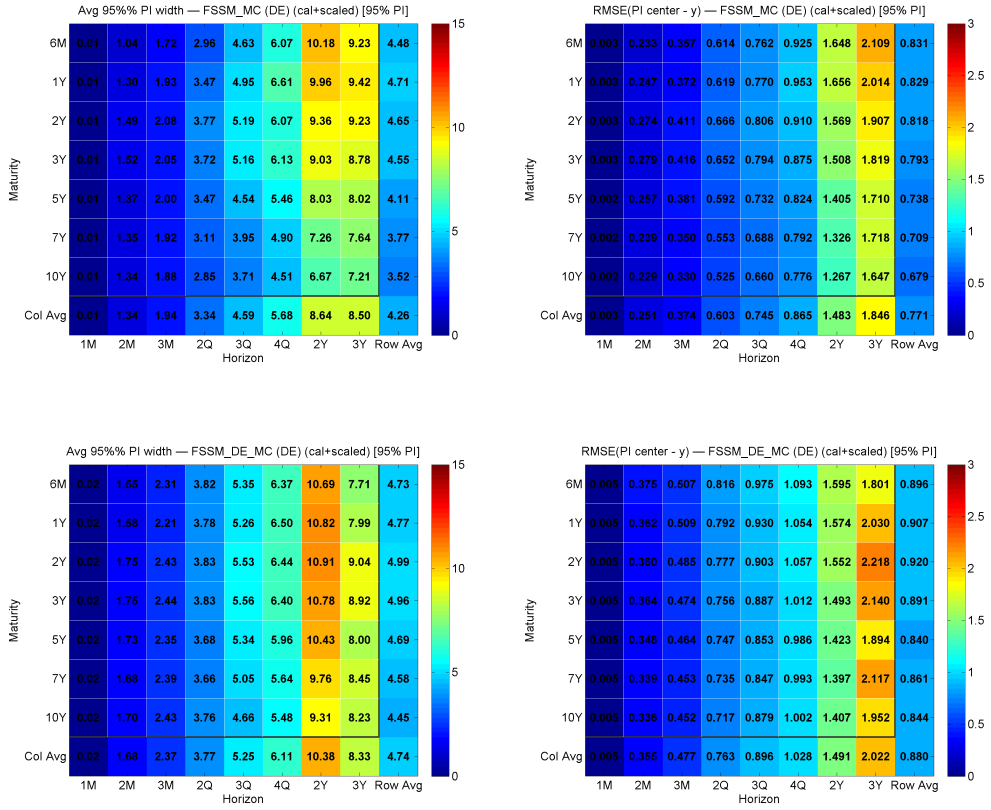
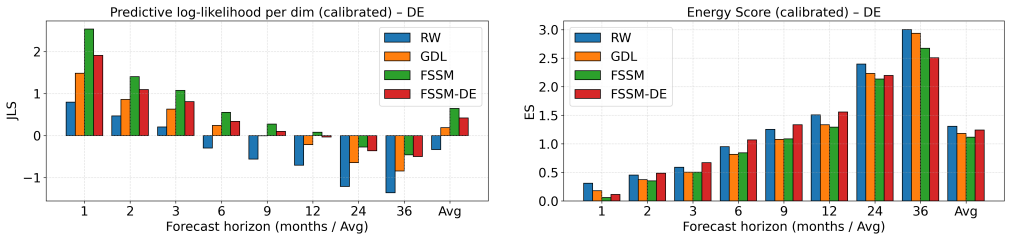


Figure 17: Monte Carlo results for country Germany and all models: Joint Log Scores and Energy Scores.



the highest JLS—making it particularly well suited for hedging and near-term risk management. Its advantage persists at intermediate horizons (2Q-3Q), where the FSSM produces the most coherent multi-horizon densities, supporting applications such as fan charts, policy briefings, and communication by the Bundesbank or the German DMO. Importantly, the model remains robust at longer horizons (2-3 years), providing realistic and internally consistent distributions that capture long-run global spillovers and policy-cycle dynamics. This makes the FSSM especially valuable for macro-policy scenario design and issuance strategy, where accurate long-horizon uncertainty is essential.

The FSSM-DE variant—where U.S. parameters are restricted to zero—shows a clear deterioration in density forecasting performance, confirming that German yields cannot be modeled in isolation. This decline is fully consistent with the well-documented pattern that ECB rate cycles lag the Fed, implying that U.S. factors contain essential forward-looking information for the German term structure.

7.3. Summary

Across the U.S. and Germany, three patterns emerge. Firstly, FSSM delivers the sharpest, most coherent, and most accurate density forecasts. It achieves the highest JLS and the lowest ES, indicating the best match to the full realized joint distribution of yields. GDL produces reasonably sharp densities with competitive marginal performance but suffers from weaker cross-equation dynamics, leading to lower JLS and ES than FSSM. RW generates the widest predictive distributions, lowest JLS, and worst ES, reflecting the absence of structural dynamics. Its densities deteriorate rapidly beyond 1-2 quarters.

In applications where the full shape of the predictive distribution matters (macro policy, scenario generation, ALM, and trading), FSSM is clearly preferred. Across both countries, and across all metrics, FSSM is the strongest density-forecasting model. Its superior JLS reflect an accurate modeling of the entire joint distribution of future yields—a critical feature for monetary-policy analysis, scenario design, DMO planning, ALM, and fixed-income strategies. The RW benchmark remains useful as a diagnostic and extremely short-horizon baseline, while GDL provides a robust alternative with respectable sharpness but weaker cross-maturity coherence. The structural FSSM model, however, consistently delivers the most credible, informative, and economically meaningful density forecasts.

8. Summary and Concluding Remarks

This paper developed a structured, data-driven state-space framework for jointly modelling and forecasting the U.S. and German yield curves. The approach embeds the empirically observed dynamic properties of international yield curve

factors—unit-root behavior in levels and curvatures, cointegration within and across countries, VAR dynamics in slopes, and rich cross-country transmission—into a unified VAR-VEC state-space representation. Within this framework, we constructed the FSSM and its several variants, which allowed us to account for patches of outliers (MShock-FSSM), investigate the predictive power of the curvature (FSSM^{LS} and MShock-FSSM^{LS}), and quantify the marginal value of foreign information for domestic prediction (FSSM-US and FSSM-DE).

Using an extended out-of-sample period (1999-2025), which covers the post-QE environment, the inflation shock of 2021-2023, the global tightening cycle, and the early normalization phase, we evaluate point forecasts, directional accuracy, and full predictive densities. Several robust findings emerge.

Point Forecasting

The RW remains difficult to beat in RMSE terms, consistent with the high persistence of yields and the dominance of high-frequency noise. The FSSM improves upon the Diebold-Li benchmarks only at medium and long horizons, where its richer dynamic structure becomes informative. The restricted variants, FSSM-US and FSSM-DE, perform worse than the full cross-country FSSM: excluding foreign information degrades predictive accuracy. This highlights the importance of modelling global linkages even when the primary objective is domestic forecasting.

Directional Accuracy

Directional forecasting provides a strikingly different picture. The RW—despite its RMSE strength—produces uninformative direction predictions. The GDL model exhibits systematic problems in anticipating turning points, owing to its indirect

treatment of country slopes and excessive smoothing in its global factor structure. By contrast, the FSSM attains substantially higher directional accuracy for both the U.S. and Germany. This improvement is economically meaningful: trading strategies based on FSSM slope signals generate large and persistent positive PnL, whereas GDL often issues wrong-direction signals and RW produces no economic value. Thus, structural modelling of slope dynamics provides information that is invisible to RMSE-based comparisons.

Density Forecasting

The most significant gains arise in full density forecasting. Using both bootstrap-based predictive distributions and Monte Carlo simulations of the state-space innovations, we evaluate models using Coverage, PI widths, ES, and JLS. Across virtually all maturities and horizons, the FSSM delivers the best calibration (coverage deviations closest to target), the most disciplined PIs, competitive multivariate ES, and the highest JLS, indicating the most coherent joint distribution of yields across maturities and horizons.

These findings hold for both the U.S. and Germany. The restricted variants again reveal the value of cross-country information: FSSM-DE performs markedly worse than the FSSM, confirming the leading role of U.S. monetary policy in shaping European yield-curve dynamics. The FSSM-US shows a deterioration relative to FSSM as well, suggesting that German information is helpful for U.S. forecasting–consistent with global spillovers.

Implications for Policy, Markets, and Risk Management

The results have direct relevance for macro-financial analysis. For short-horizon desk decisions (1-3 months), when predictability is low, simple bench-

marks such as RW or GDL suffice for point forecasts. However, FSSM remains superior for directional signals and density coherence. For medium-horizon risk management and ALM (2-6 quarters), the FSSM provides the most internally consistent predictive distributions, essential for stress scenarios, fan charts, balance-sheet risk, and regulatory capital planning. For macro-policy and issuance strategy (2-3 years), at longer horizons, the structural design of the FSSM—cointegration, cross-country dynamics, and multi-maturity transmission—delivers the most realistic tail behavior and scenario paths, providing a superior tool for central-bank communication and government debt-management offices.

Overall Assessment

Taken together, the evidence shows that while the RW remains a formidable competitor for short-run point forecasts, it conveys no forward-looking information in economic terms. The GDL model performs well in levels but struggles with direction and density coherence. The FSSM dominates in all economically relevant dimensions: directional accuracy, multivariate sharpness, calibrated predictive uncertainty, and joint distribution fit. The performance of the cross-country model relative to its country-only variants underscores the importance of global linkages.

In summary, incorporating structural cross-country dynamics is essential for reliable yield-curve prediction. The FSSM provides the most coherent and informative forecasts of levels, changes, and full predictive densities, offering substantial gains for both academic modelling and real-world monetary policy and fixed-income decision-making.

Appendix A. Appendix: Modeling Framework

The shock variables $\mathbf{\Lambda}_t^{\text{VAR}}$ and $\mathbf{\Lambda}_t^{\text{VEC}}$ in the MShock-FSSM model in Equation 10 are defined as:

$$\mathbf{\Lambda}_{t,s_{US}}^{\text{VAR}} = 0 \tag{A.1}$$

$$\mathbf{\Lambda}_{t,s_{DE}}^{\text{VAR}} = \begin{cases} 1, & t = \text{'2008:08'} \\ -1, & t = \text{'2008:09'} \\ 0, & \text{otherwise} \end{cases} \tag{A.2}$$

$$\mathbf{\Lambda}_{t,l_{US}}^{\text{VEC}} = \begin{cases} 1, & t = \text{'2008:10'} \\ -1, & t = \text{'2008:12'} \\ 0, & \text{otherwise} \end{cases} \tag{A.3}$$

$$\mathbf{\Lambda}_{t,l_{DE}}^{\text{VEC}} = \begin{cases} 1, & t = \text{'2008:10'} \\ 0, & \text{otherwise} \end{cases} \tag{A.4}$$

$$\mathbf{\Lambda}_{t,c_{US}}^{\text{VEC}} = \begin{cases} -1, & t = \text{'2008:10'} \\ 0, & \text{otherwise} \end{cases} \tag{A.5}$$

$$\mathbf{\Lambda}_{t,c_{DE}}^{\text{VEC}} = \begin{cases} -1, & t = \text{'2008:10'} \\ 0, & \text{otherwise} \end{cases} \tag{A.6}$$

Appendix B. Directional Accuracy

Appendix B.1. Directional Forecast Metric

Let y_t denote the variable of interest (yield, factor, or slope) observed at time t, and let $\hat{y}_{t+h|t}$ be the h-step-ahead forecast made at time t. We define the predicted direction of change as

$$d_{t,h}^{\text{pred}} = \text{sign}\left(\hat{y}_{t+h|t} - y_t\right),$$

and the realized direction as

$$d_{t,h}^{\text{real}} = \text{sign}\left(y_{t+h} - y_t\right),$$

where $\text{sign}(\cdot)$ is the usual sign function.

A directional "hit" occurs whenever the sign of the forecasted change coincides with the sign of the realized change. We record this with the indicator

$$\mathbb{1}_{t,h} = \begin{cases} 1, & \text{if } d_{t,h}^{\text{pred}} = d_{t,h}^{\text{real}}, \\ 0, & \text{otherwise.} \end{cases}$$

For a given horizon h, and with T_h out-of-sample forecasts available, the directional accuracy (DA) of a model is

$$\text{DA}(h) = \frac{1}{T_h} \sum_t \mathbb{1}_{t,h}.$$

A value of $\text{DA}(h) = 0.5$ corresponds to chance performance (a fair coin), whereas values significantly above 0.5 indicate economically relevant directional information.

Under the RW benchmark,

$$\hat{y}_{t+h|t}^{\text{RW}} = y_t \quad \Rightarrow \quad \text{sign}\left(\hat{y}_{t+h|t}^{\text{RW}} - y_t\right) = 0,$$

so the RW is effectively directionally uninformative; its DA is 0 (as it predicts no change) and it provides no exploitable trading signal.

Appendix B.2. From Directional Signals to Trading Profits

To connect directional accuracy to economic value, we map forecasts of the slope factor s_t (10Y–2Y) into a stylized trading strategy.

For each model (in our case, RW, GDL, and FSSM) and horizon h, we construct a trading signal

$$\text{Signal}_{t,h} = \text{sign}\left(\hat{s}_{t+h|t} - s_t\right) \in \{-1, 0, +1\},$$

where $+1$ denotes a *steepener* position (slope increases) and -1 a *flattener* (slope decreases). A value of 0 corresponds to no position.

The realized change in the slope over the horizon h is

$$\Delta s_{t,h} = s_{t+h} - s_t.$$

The (per-unit-notional) profit-and-loss (PnL) generated by the model at horizon h is then

$$\text{PnL}_{t,h} = \text{Signal}_{t,h} \cdot \Delta s_{t,h}.$$

Cumulative PnL over the out-of-sample period at horizon h is

$$\text{CumPnL}(h) = \sum_t \text{PnL}_{t,h}.$$

Because the strategy uses only the *sign* of the model forecast, these PnL measures are driven by directional information rather than point-forecast precision.

Appendix B.3. Why FSSM Excels and GDL Underperforms in Direction

The superior directional performance of the FSSM is a direct consequence of its structural design. The U.S. and German slopes enter the state vector directly and evolve according to a bivariate VAR(5). This specification captures persistence,

cross-country interactions, and turning points in slope dynamics. Because slopes are genuine state variables, the model is forward-looking in the evolution of the yield curve shape. Furthermore, the level–curvature block follows a VEC specification that enforces cointegration between the U.S. and German factors. These error-correction terms are crucial for predicting medium- and long-horizon directional shifts that RW and GDL cannot capture. In addition, slope reversals are often preceded by joint movements in levels and curvatures and by cross-country lead–lag patterns. By modelling all drivers jointly, the FSSM detects these reversals earlier than GDL, leading to correct directional calls where GDL tends to extrapolate past trends.

The GDL model, by contrast, suffers from two structural limitations. Firstly, GDL forecasts global factors and idiosyncratic components and then reconstructs country-specific slopes by refitting a Nelson–Siegel cross-section to the forecasted yields. Slopes are thus a by-product of yield forecasts, not governed by their own dynamics. Small rotations in the refitted cross-section can easily flip the implied slope sign, even when the underlying yield forecasts are accurate. Secondly, the combination of PCA extraction, static projections from global to local factors, VAR(1) dynamics, and Nelson–Siegel refitting generates very smooth slope paths. After a prolonged flattening episode, GDL tends to extrapolate further flattening and therefore misses the onset of steepening, issuing wrong-signed signals exactly when directional information is most valuable.

Appendix C. Density Forecasting: Bootstrapping Method

Appendix C.1. Notation and Forecast Errors

We index forecast origins by $t \in \mathcal{T} = \{1, \dots, T\}$, horizons by $h \in \mathcal{H} = \{1, \dots, H\}$, and maturities by $\tau \in J = \{1, \dots, J\}$.

For each model, the point forecast of the m-maturity yield issued at origin t for horizon h is denoted

$$\hat{y}_{t,h,\tau},$$

and the aligned realization is

$$y_{t,h,\tau},$$

so that the h-step forecast error is

$$e_{t,h,\tau} = y_{t,h,\tau} - \hat{y}_{t,h,\tau}.$$

For a given (t, h) we collect the J-dimensional yield curve and its error as

$$\hat{\mathbf{y}}_{t,h} = (\hat{y}_{t,h,1}, \dots, \hat{y}_{t,h,J})', \quad \mathbf{y}_{t,h} = (y_{t,h,1}, \dots, y_{t,h,J})', \quad \mathbf{e}_{t,h} = \mathbf{y}_{t,h} - \hat{\mathbf{y}}_{t,h}.$$

Appendix C.2. Residual Bootstrap Pipeline

For each model we apply the *same* residual bootstrap pipeline. The goal is to obtain, at each (t, h), a set of joint predictive draws $\{\mathbf{X}_{t,h}^{(b)}\}_{b=1}^{B}$ for the full yield curve, and then to calibrate those draws so that marginal coverage and dispersion are comparable across models. More specifically, the steps we make are as follows:

1. *Robust centering and studentization.* In this first step, we mitigate for the presence of outliers (shocks) by centering the forecast errors (using robust location estimators) and scaling them to remove level shifts and time-varying volatility that would otherwise distort the bootstrap.

2. *Vector moving block bootstrap.* We apply a multivariate moving block bootstrap (MBB) to the error vectors $\mathbf{e}_{t,h}$. Blocks are resampled as *vectors* so that temporal dependence and cross-maturity dependence are preserved.

3. *Construction of raw predictive draws.* For each (t, h) we add resampled error blocks to the point forecast $\hat{\mathbf{y}}_{t,h}$ to obtain bootstrap draws $\{\mathbf{X}_{t,h}^{(b)}\}_{b=1}^{B}$ and corresponding marginal quantiles $q_{t,h,\tau}^{\text{lo}}, q_{t,h,\tau}^{\text{hi}}$.

4. *PIT calibration (marginal calibration).* Probability integral transforms for each maturity are used to estimate nonparametric calibration maps that correct residual under- or over-coverage of the raw predictive distributions.

5. *Horizon-maturity specific variance rescaling.* For each (h, τ) we learn a multiplicative scale factor that adjusts the predictive variance without altering the median, so that the empirical hit rate of the 90%, 95%, and 99% bands moves toward the nominal level.

6. *Final calibrated draws and scoring.* In this final step, the bootstrap is rerun using the PIT calibration maps and scale factors to generate the final calibrated draws and prediction intervals. All density metrics[23] (Coverage, Average Width of Prediction Intervals (PIs), RMSE of PIs, Joint Log Scores, and Energy Scores) are computed from this calibrated output, ensuring a fair comparison across models.

Appendix C.3. Evaluation Metrics

The metrics we use for the evaluation of the density forecasting performance are summarized in Table C.5. In addition, we calculate the RMSE of the PIs centers

[23] A more detailed explanation of these metrics is provided in the Appendix Appendix C.

as follows. For each (t, h, τ) we define the center of the 90%, 95%, and 99% PIs as

$$c_{t,h,\tau} = \frac{q^{\text{lo}}_{t,h,\tau} + q^{\text{hi}}_{t,h,\tau}}{2}.$$

This midpoint serves as an estimate of the location of the predictive distribution (similar to the mean or median when the distribution is approximately symmetric). For each triplet (t, h, τ) for which all quantities are available, the forecast error is

$$e^{\text{PI}}_{t,h,\tau} = c_{t,h,\tau} - y_{t,h,\tau}.$$

For each horizon and maturity we compute

$$\text{RMSE}_{h,\tau} = \sqrt{\frac{1}{|\mathcal{T}_{h,\tau}|} \sum_{t \in \mathcal{T}_{h,\tau}} \left(e^{\text{PI}}_{t,h,\tau}\right)^2}.$$

The collection of values $\{\text{RMSE}_{h,\tau}\}_{h,\tau}$ forms a $H \times J$ heatmap that summarizes how well-centered the predictive distributions are at each forecasting horizon and maturity. The metric evaluates the point-forecast accuracy *implied by* the PIs. Small values of $\text{RMSE}_{h,\tau}$ indicate that the midpoint of the PI is typically close to the realized yield for that horizon and maturity. Large values signal either systematic bias (interval centers consistently too high or too low) or excessive dispersion in the predictive location.

Appendix C.4. Additional Results: Coverage, 90%, 95%, 99%

Appendix C.5. Additional Results: PI Widths and RMSE of PI Widths, 90%, 99%

Table C.5: Evaluation metrics for density forecasting.

Metric	Definition	Purpose	Key Insights
Coverage (95% PI)	Share of realizations inside the 95% prediction interval.	Check marginal *calibration* (hit rate).	**Target** ≈ 0.95. Below $\Rightarrow$ under-coverage (too tight); above $\Rightarrow$ over-coverage (too wide). Inspect by horizon and maturity.
Average PI Width	Mean distance between upper and lower prediction limits.	Measure *sharpness* once coverage is okay.	**Lower is better only if coverage is near target**; otherwise narrow bands can be misleading. Compare within same country/maturity set.
Energy Score (ES)	Multivariate distance-based score from draws to the realization (and among draws).	Assess joint accuracy of center + spread without a distributional assumption.	**Lower is better**. Penalizes bias and mis-dispersion; *scale-dependent* (compare like with like). Improvements most meaningful when coverage is acceptable.
Joint Log Score / dim (JLS, MVN)	Log predictive density of the realized curve under an MVN fit to the draws, divided by # maturities.	Test joint fit of mean and *covariance* (co-movement).	**Higher is better**. Sensitive to centering and dependence; stabilized via shrinkage/ridge; can penalize heavy-tailed data under MVN. Use alongside ES for a balanced view.

Figure C.18: Boostrapping results for RW and GDL and country U.S.: Coverage at 90% and 95%

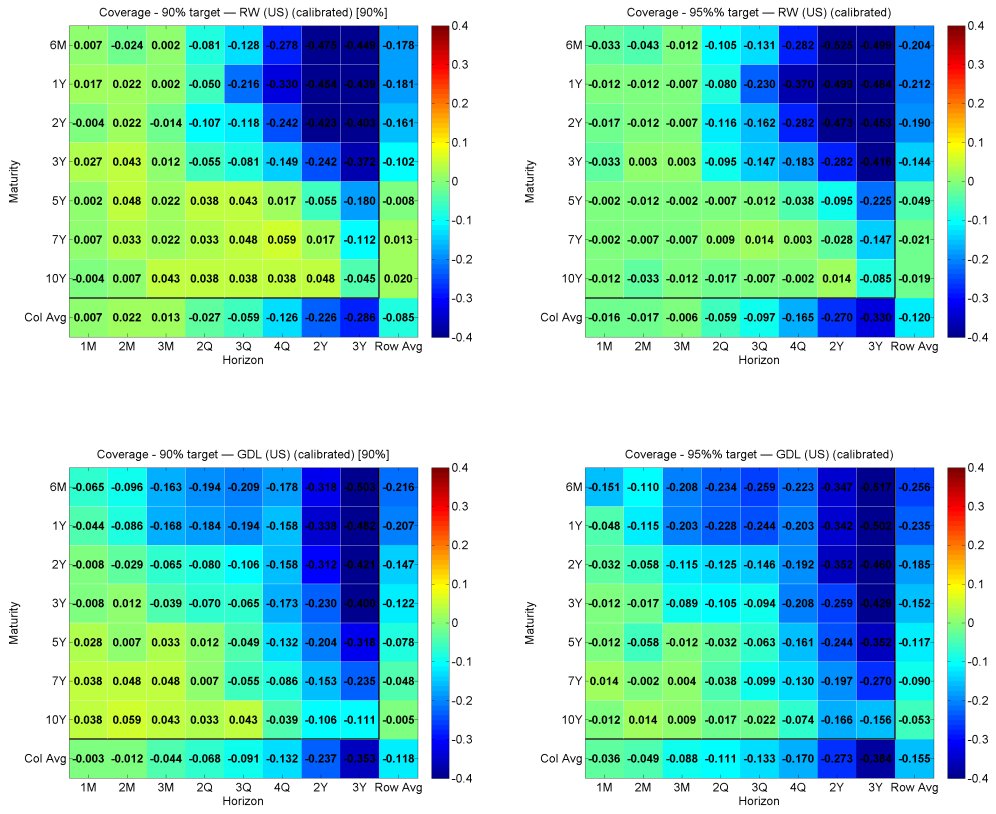


Figure C.19: Boostrapping results for RW and GDL and country U.S.: Coverage at 99%

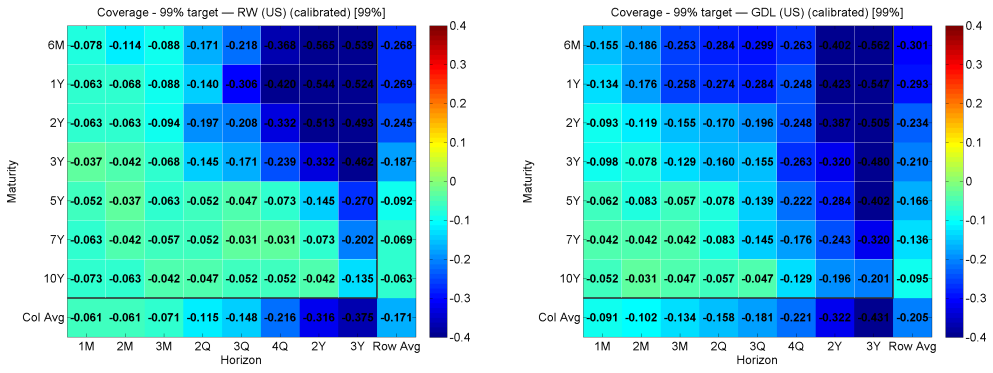


Figure C.20: Boostrapping results for FSSM and FSSM-US and country U.S.: Coverage at 90% and 95%

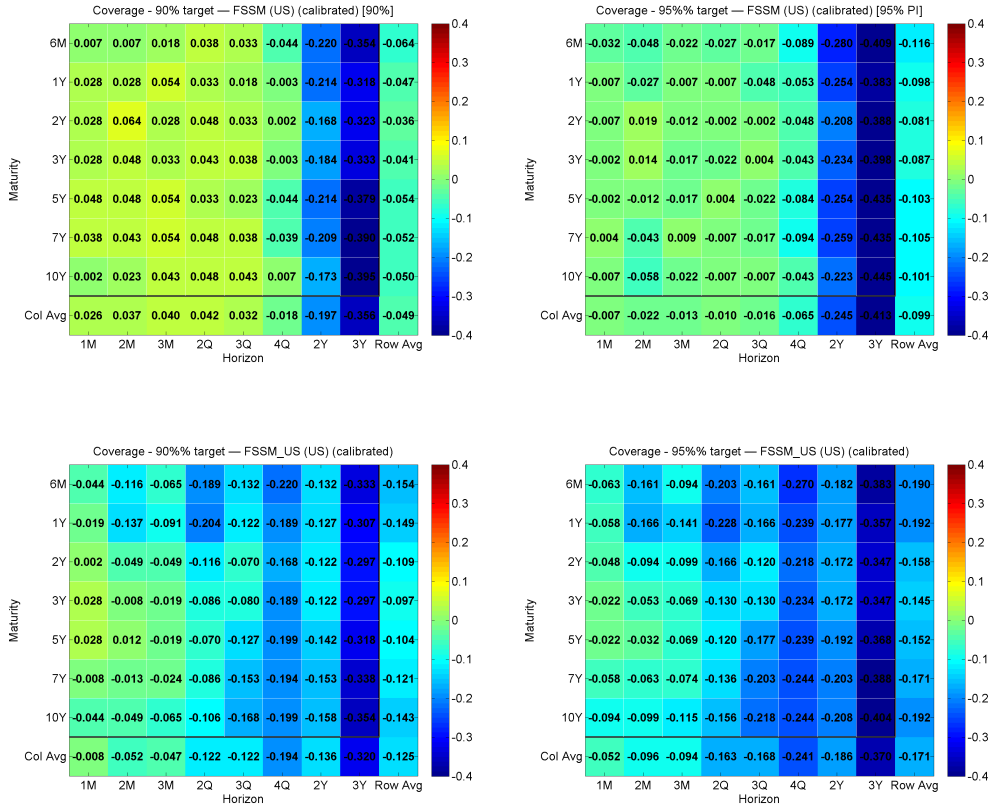


Figure C.21: Boostrapping results for FSSM and FSSM-US and country U.S.: Coverage at 99%

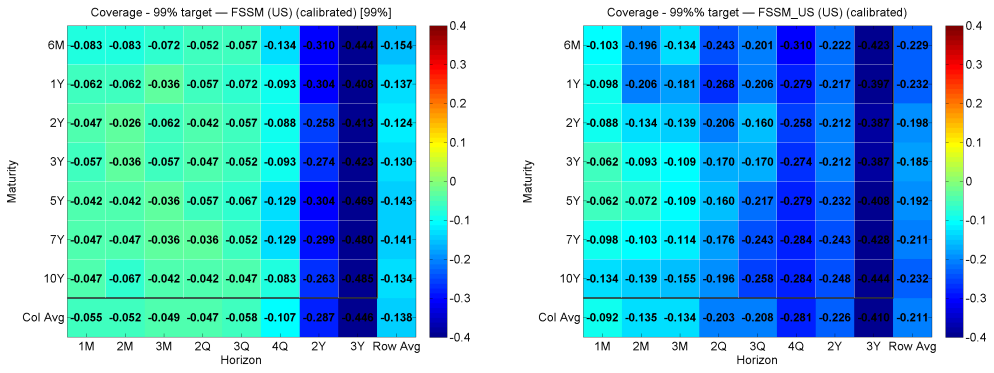


Figure C.22: Boostrapping results for RW and GDL and country Germany: Coverage at 90% and 95%

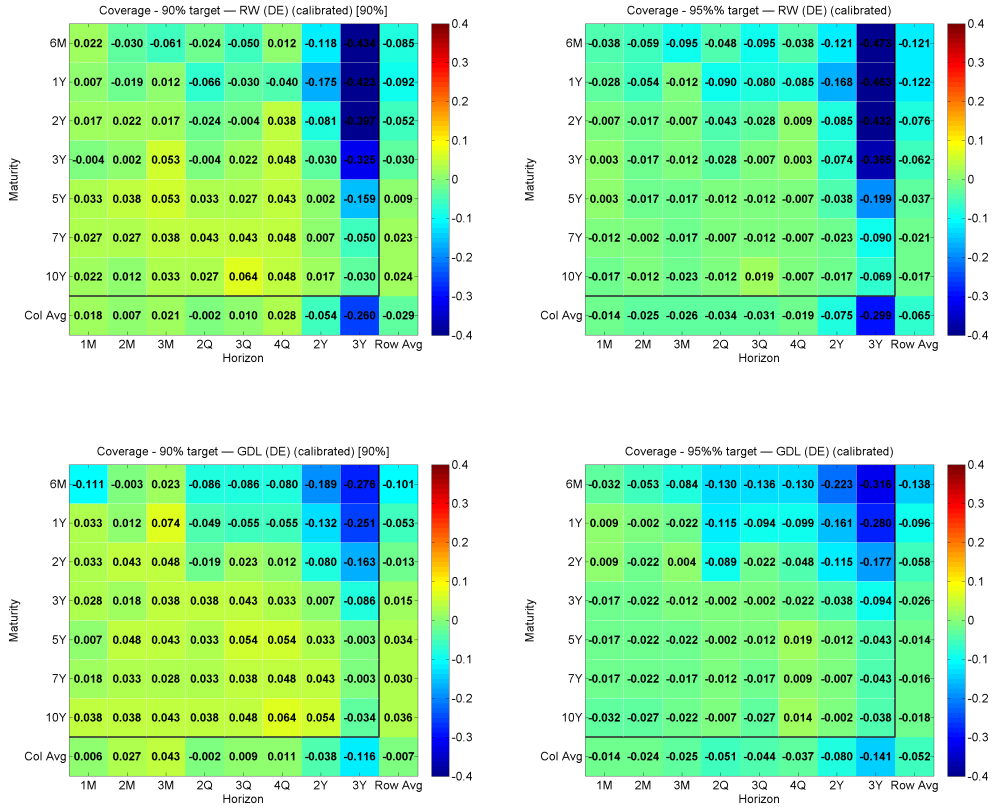


Figure C.23: Boostrapping results for RW and GDL and country Germany: Coverage at 99%

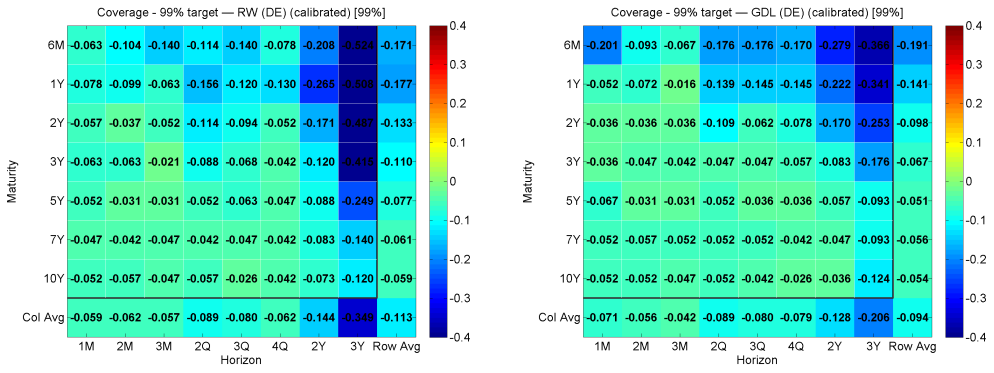


Figure C.24: Boostrapping results for FSSM and FSSM-DE and country Germany: Coverage at 90% and 95%

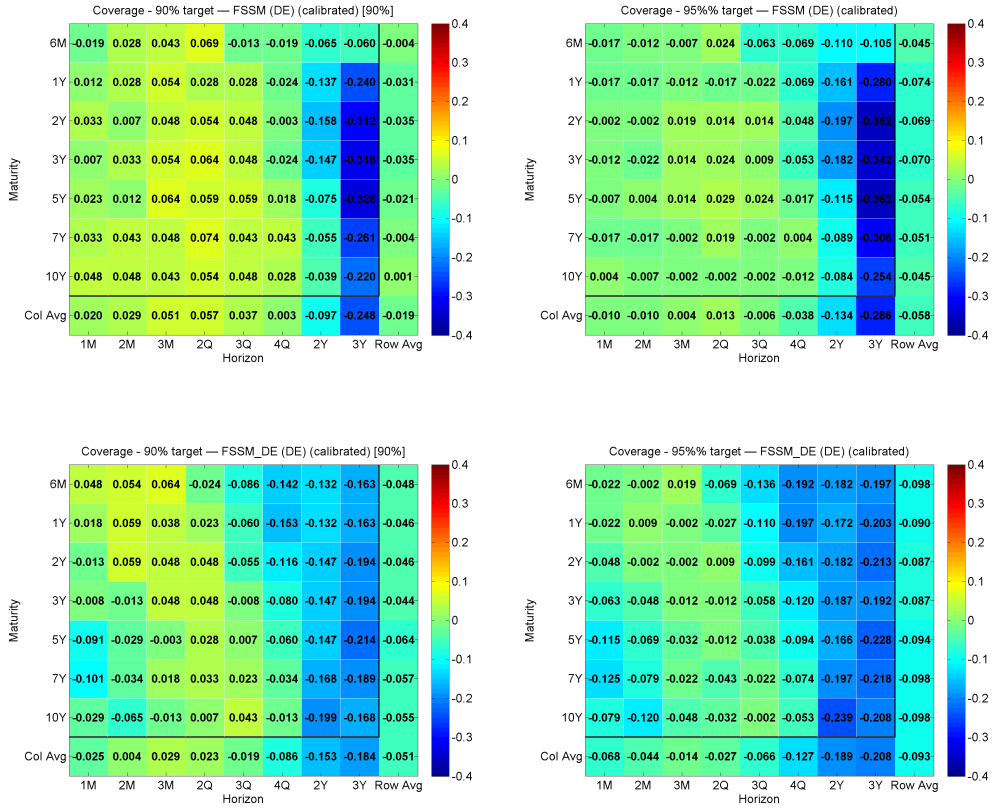


Figure C.25: Boostrapping results for FSSM and FSSM-DE and country Germany: Coverage at 99%

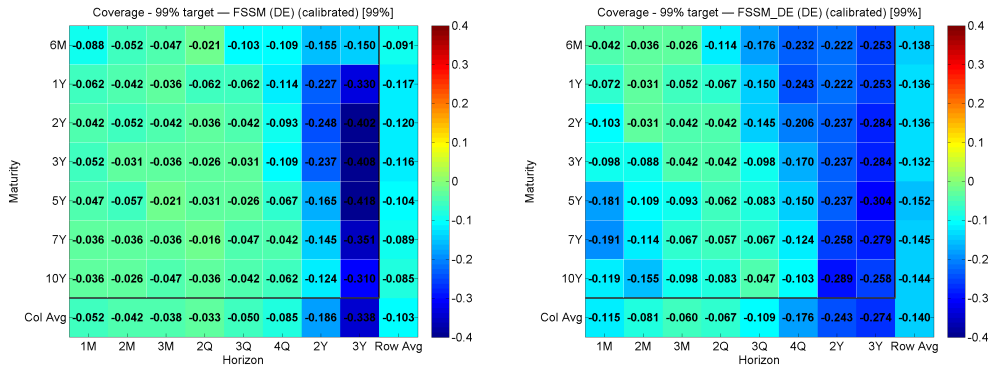


Figure C.26: Boostrapping results for RW and GDL and country U.S.: Average PI Widths and their RMSE at 90%

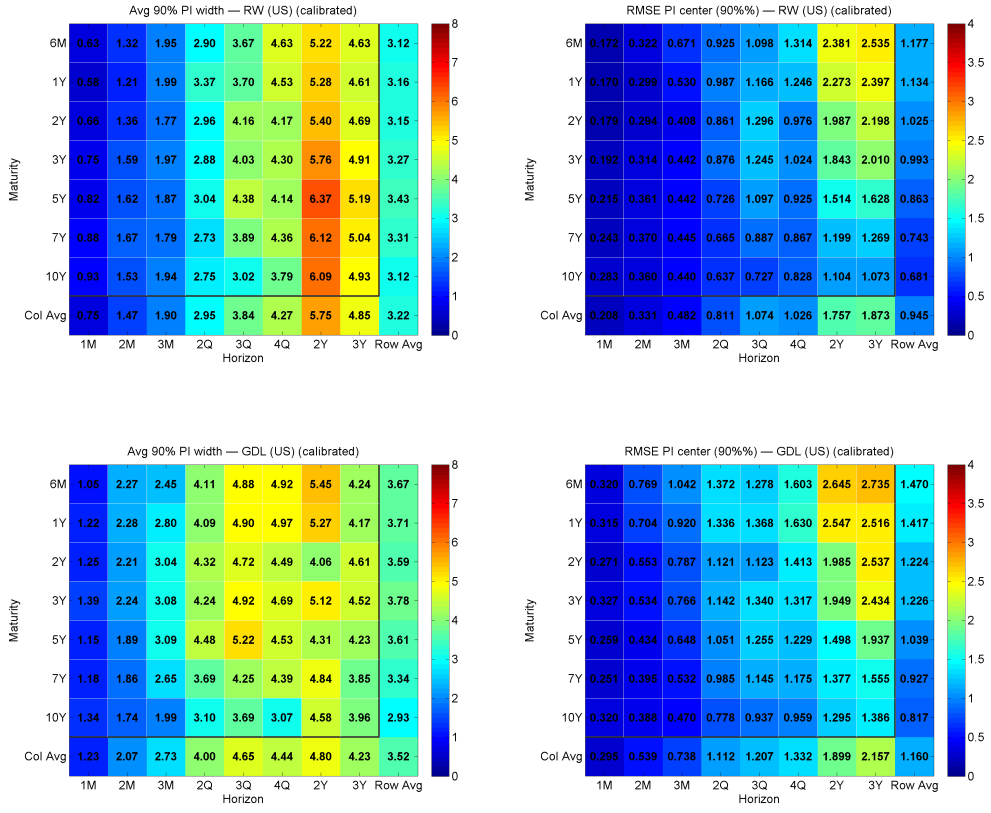


Figure C.27: Boostrapping results for FSSM and FSSM-US and country U.S.: Average PI Widths and their RMSE at 90%

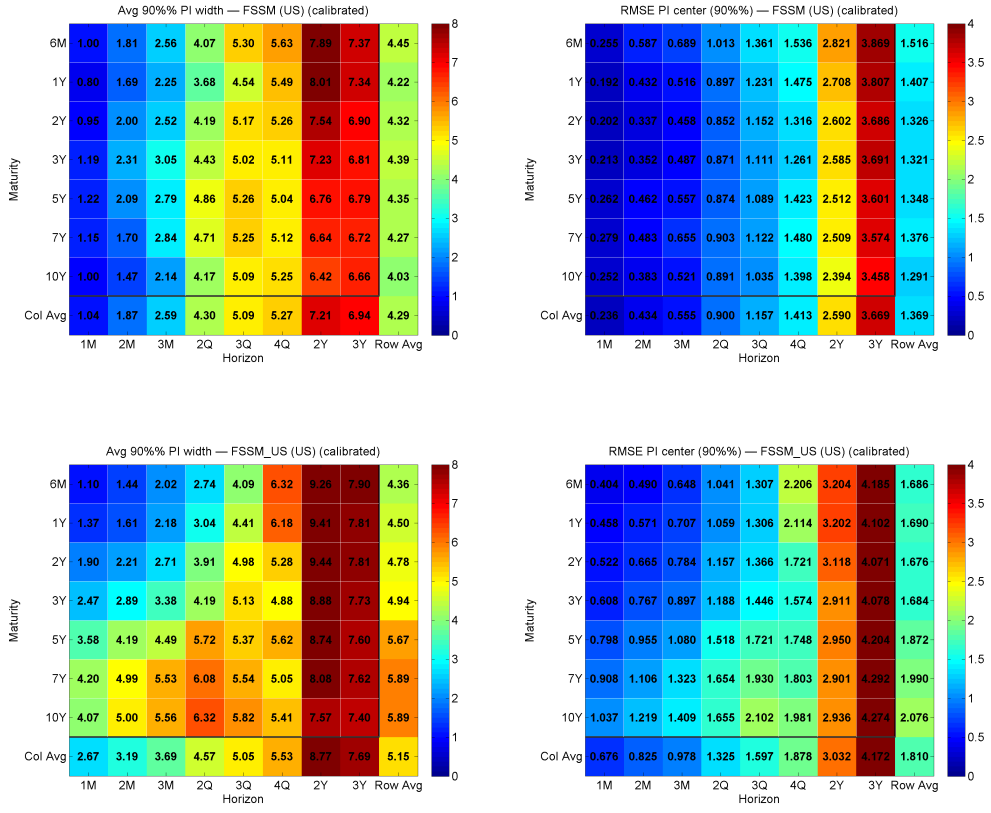


Figure C.28: Boostrapping results for RW and GDL and country Germany: Average PI Widths and their RMSE at 90%

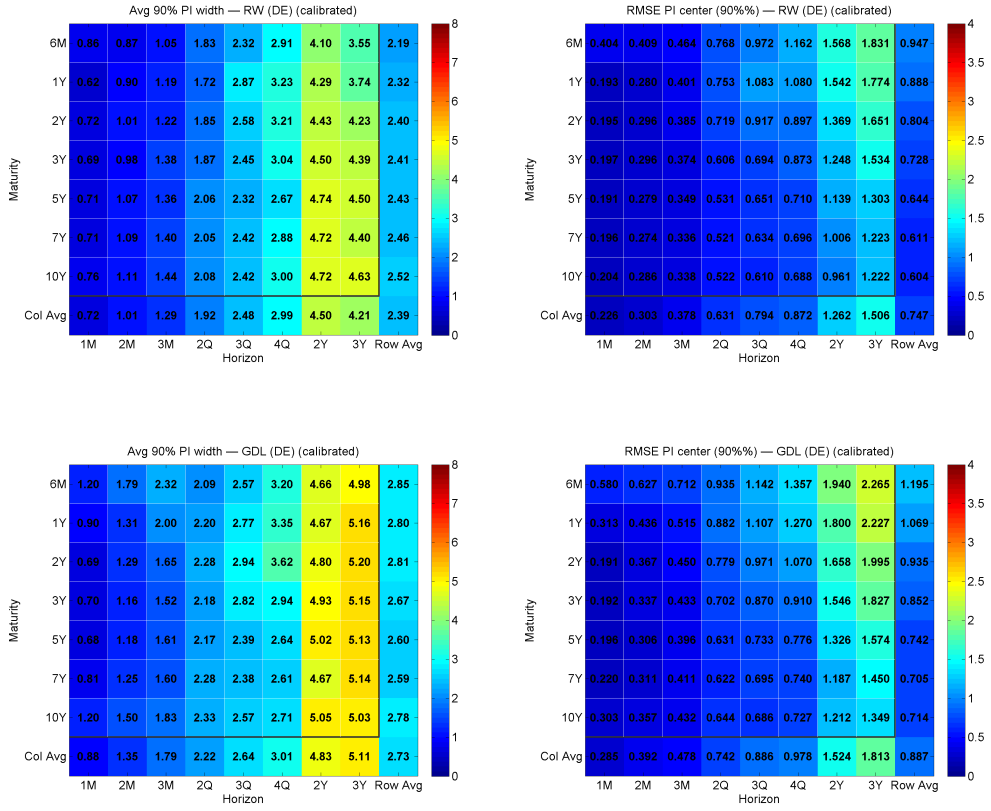


Figure C.29: Boostrapping results for FSSM and FSSM-DE and country Germany: Average PI Widths and their RMSE at 90%

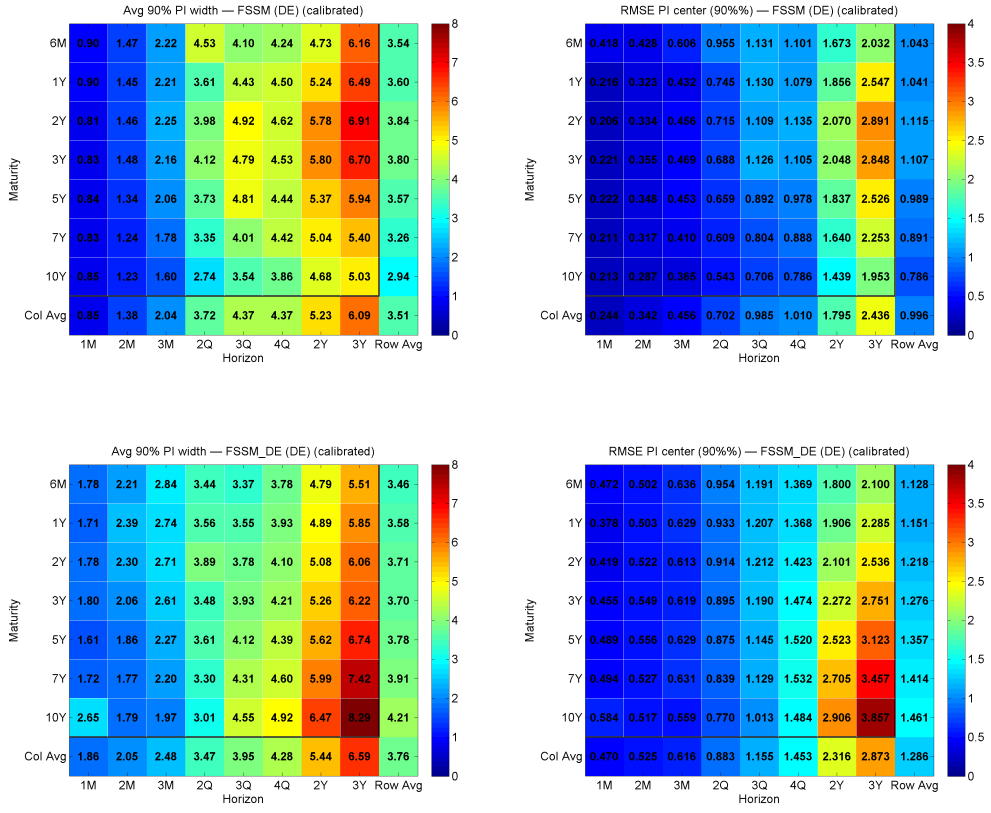


Figure C.30: Boostrapping results for RW and GDL and country U.S.: Average PI Widths and their RMSE at 99%

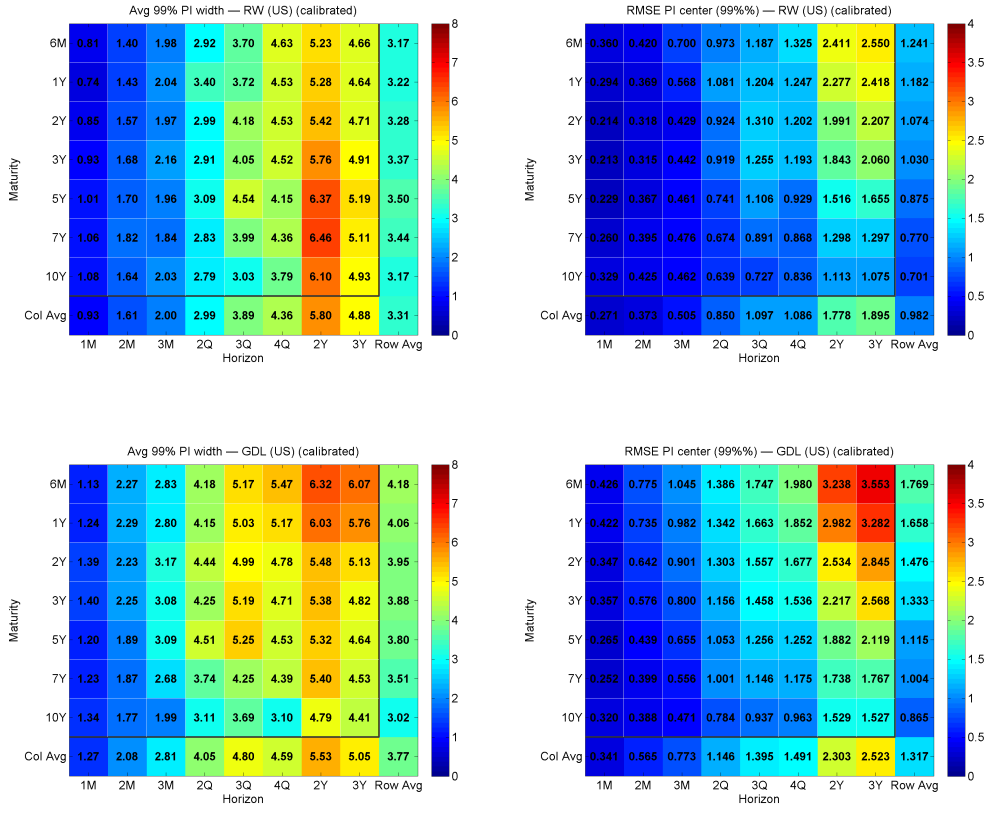


Figure C.31: Boostrapping results for FSSM and FSSM-US and country U.S.: Average PI Widths and their RMSE at 99%

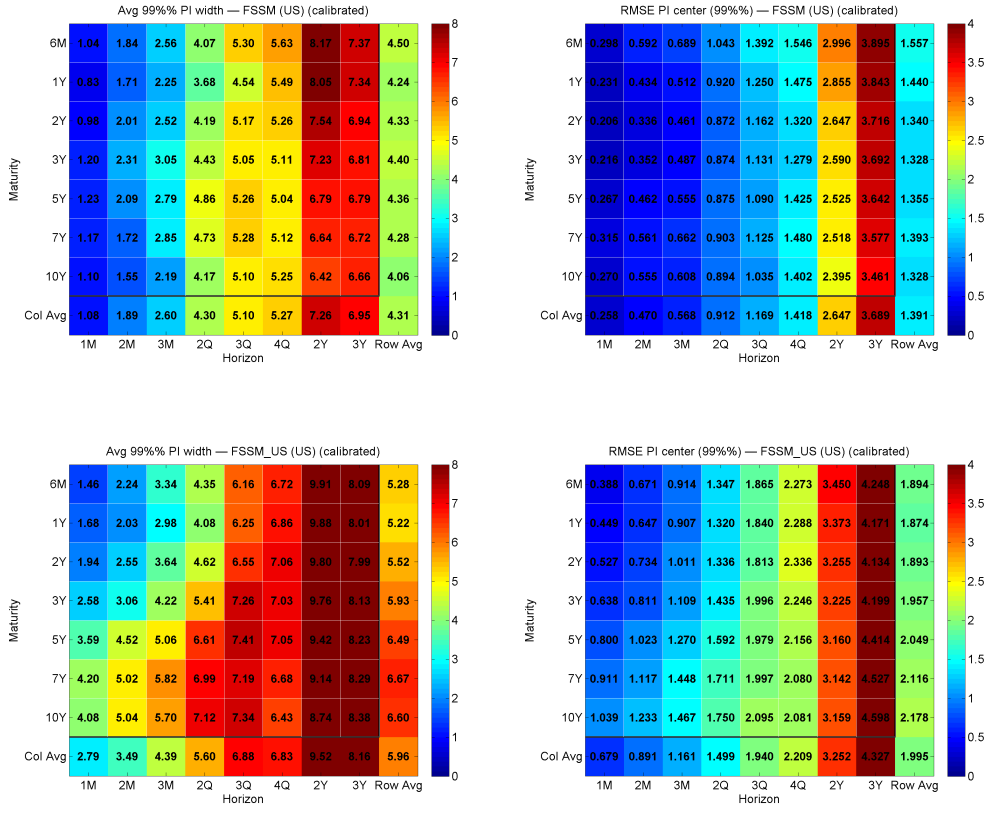


Figure C.32: Boostrapping results for RW and GDL and country Germany: Average PI Widths and their RMSE at 99%

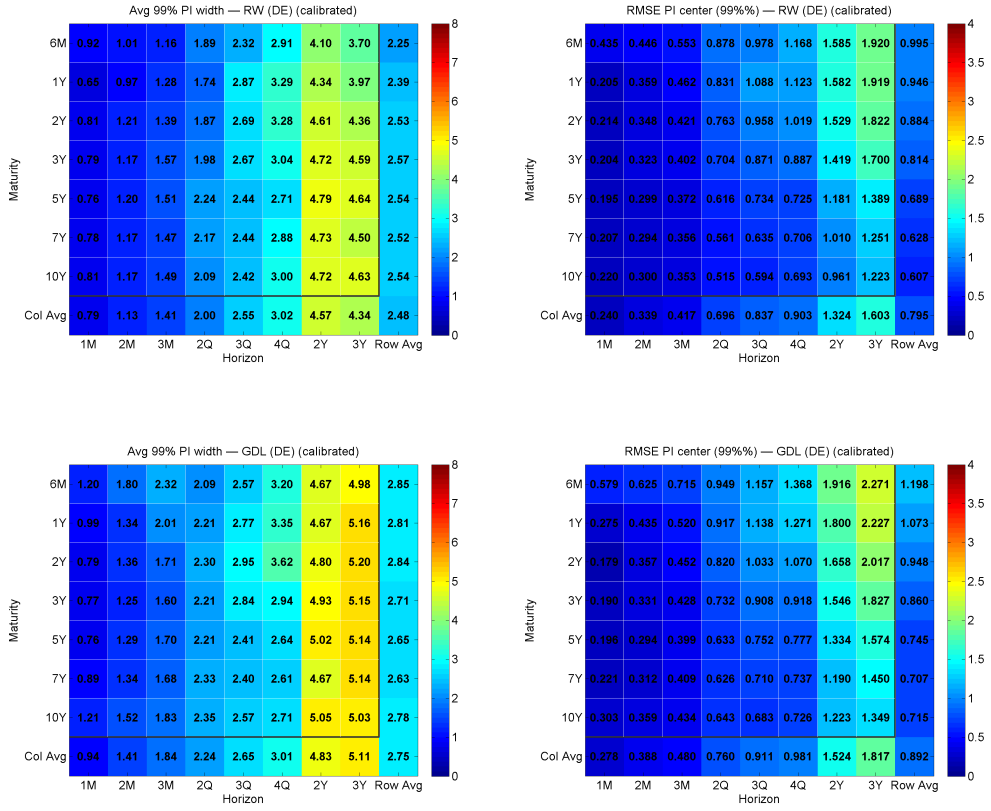
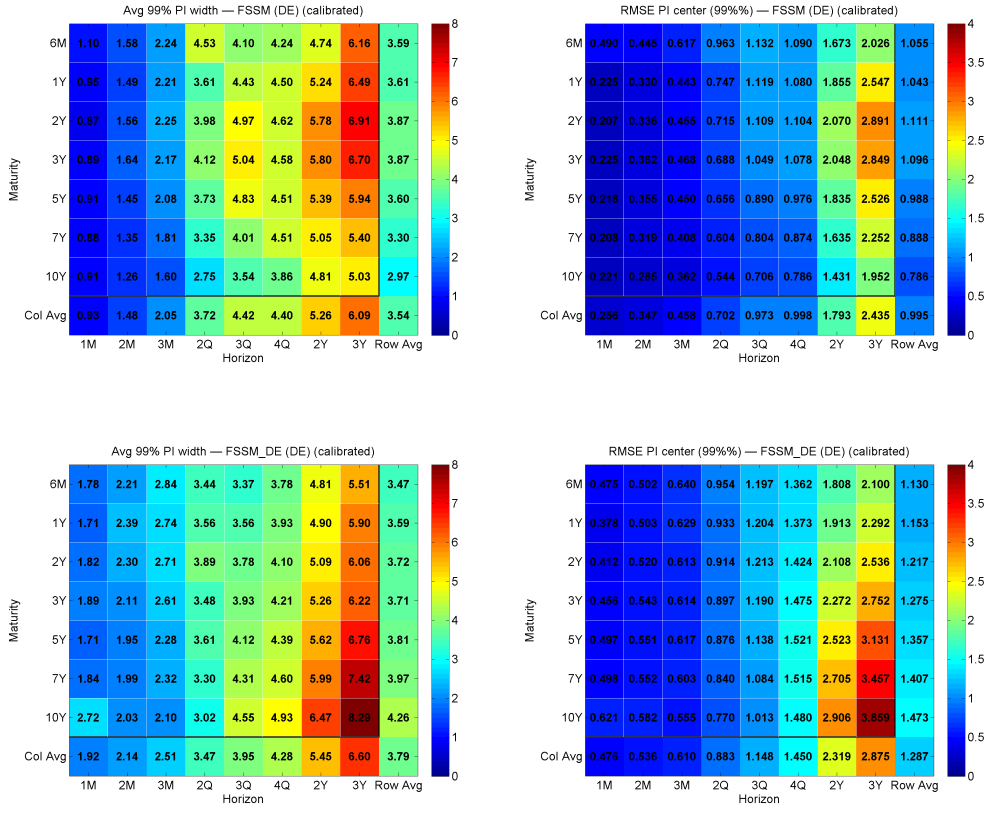


Figure C.33: Boostrapping results for FSSM and FSSM-US and country Germany: Average PI Widths and their RMSE at 99%



Appendix D. Density Forecasting: Monte Carlo Method

Appendix D.1. Monte Carlo Pipeline

For each forecast origin t and horizon h, and for each Monte Carlo (MC) draw $b = 1, \ldots, B$ (with $B = 2000$):

1. We simulate state shocks $\boldsymbol{\epsilon}_t$ from the transition equation (8) and measurement shocks $\boldsymbol{u}_t, \boldsymbol{\eta}_t$ from the observation equation (9).

2. Using these innovations, we recursively propagate the latent factors $\boldsymbol{x}_t$ and $\Delta\boldsymbol{x}_t$ forward in time, obtaining a *joint simulated factor distribution* for all horizons.

3. For each simulated factor path, we map factors into yields via the Diebold–Li formula, producing a full set of predictive yield draws $\{\mathbf{X}_{t,h}^{(b)}\}_{b=1}^{B}$ across maturities.

4. The predictive draws are then processed identically across all models: empirical quantiles, PIT calibration, horizon–maturity-specific variance scaling, and final calibrated predictive intervals.

5. We compute the scoring metrics summarized in Section 6 and Appendix C.

All post-simulation steps are kept identical across models. Thus, differences in performance reflect only the models' internal dynamics.

Appendix D.2. Additional Results: Coverage, 90%, 95%, 99%

Appendix D.3. Additional Results: PI Widths and RMSE of PI Widths, 90%, 99%

Figure D.34: Monte Carlo results for RW and GDL and country U.S.: Coverage at 90% and 95%

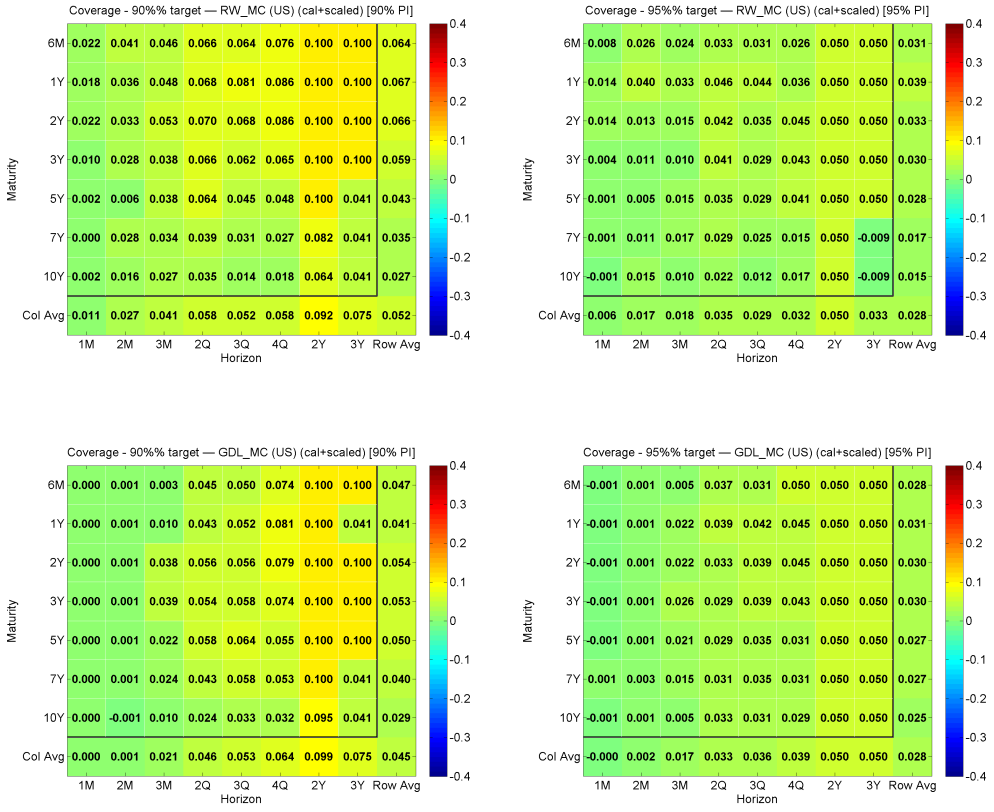


Figure D.35: Monte Carlo results for RW and GDL and country U.S.: Coverage at 99%

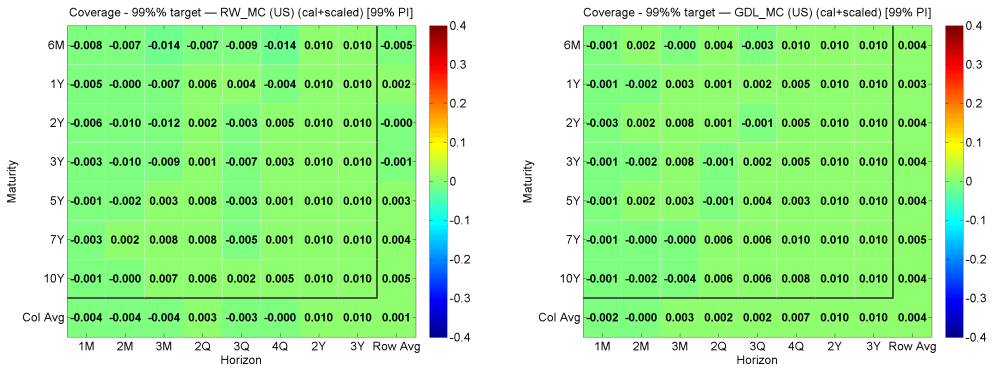


Figure D.36: Monte Carlo results for FSSM and FSSM-US and country U.S.: Coverage at 90% and 95%

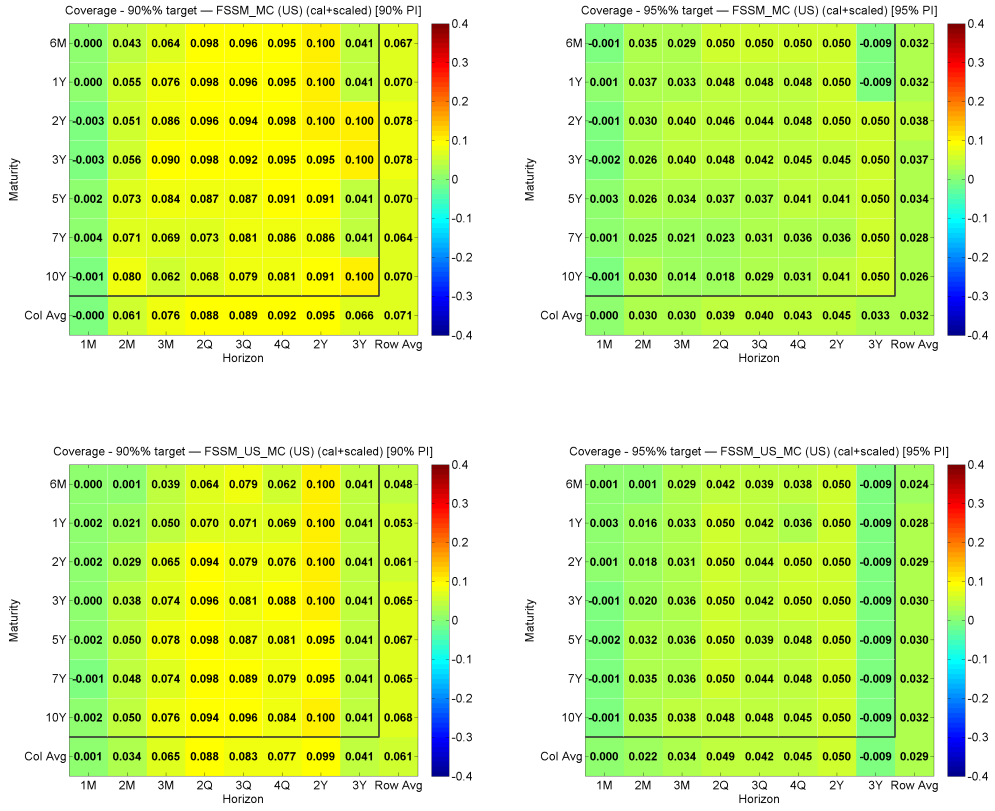


Figure D.37: Monte Carlo results for FSSM and FSSM-US and country U.S.: Coverage at 99%

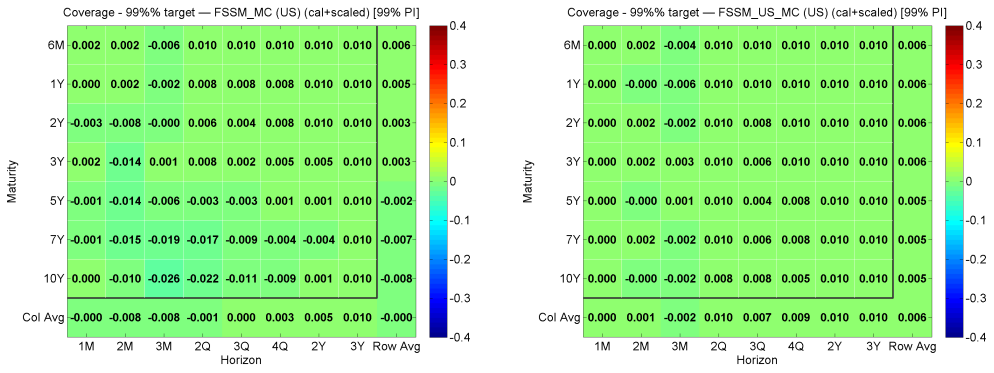


Figure D.38: Monte Carlo results for RW and GDL and country Germany: Coverage at 90% and 95%

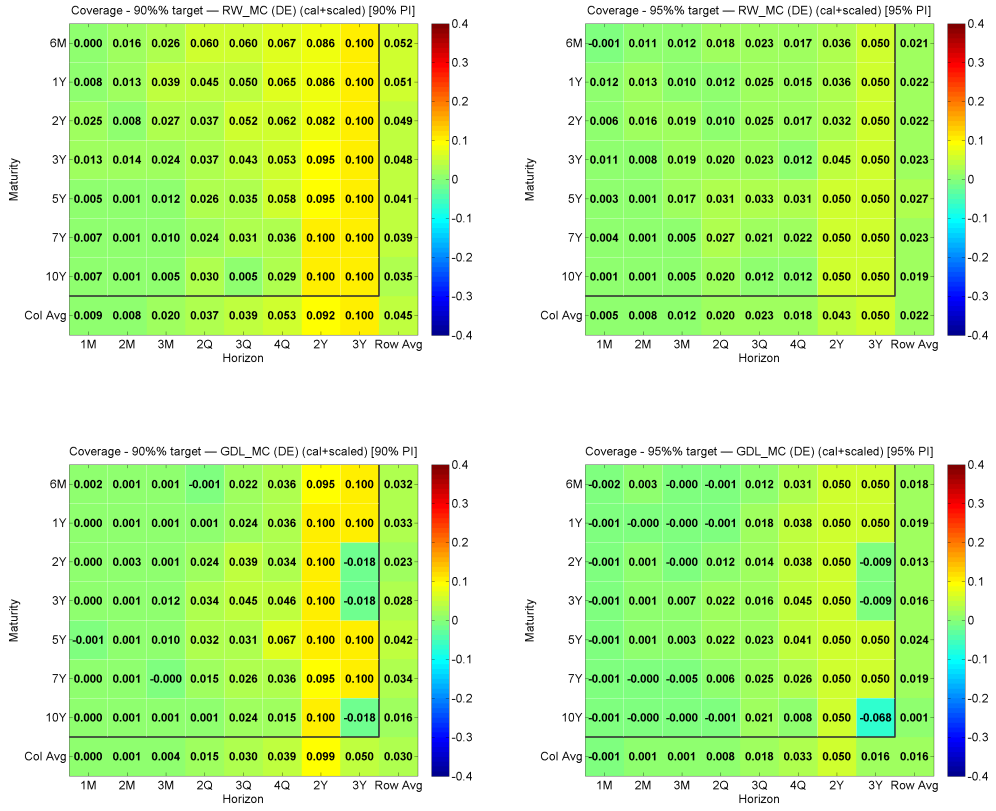


Figure D.39: Monte Carlo results for RW and GDL and country Germany: Coverage at 99%

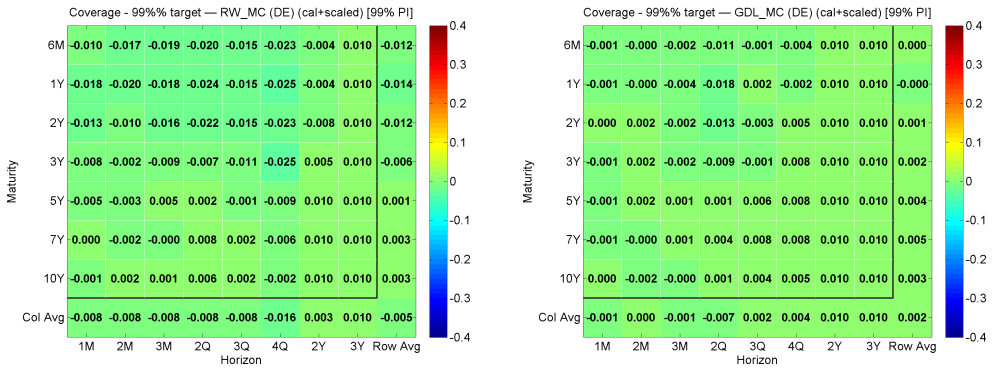


Figure D.40: Monte Carlo results for FSSM and FSSM-DE and country Germany: Coverage at 90% and 95%

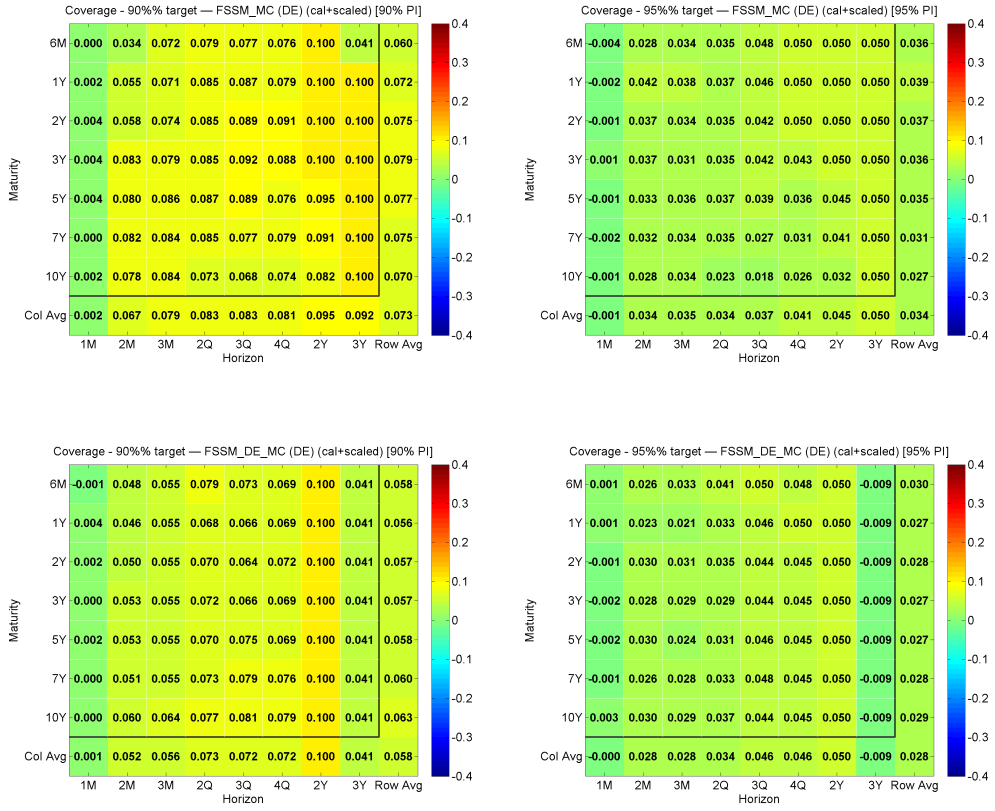


Figure D.41: Monte Carlo results for FSSM and FSSM-DE and country Germany: Coverage at 99%

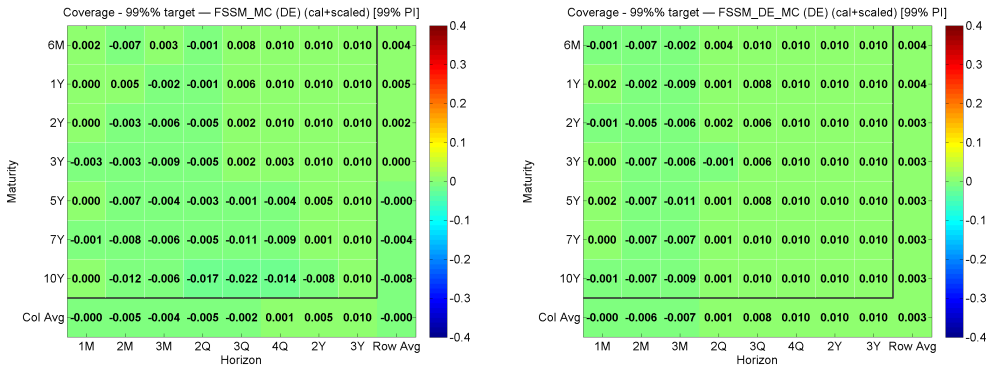


Figure D.42: Monte Carlo results for RW and GDL and country U.S.: Average PI Widths and their RMSE at 90%

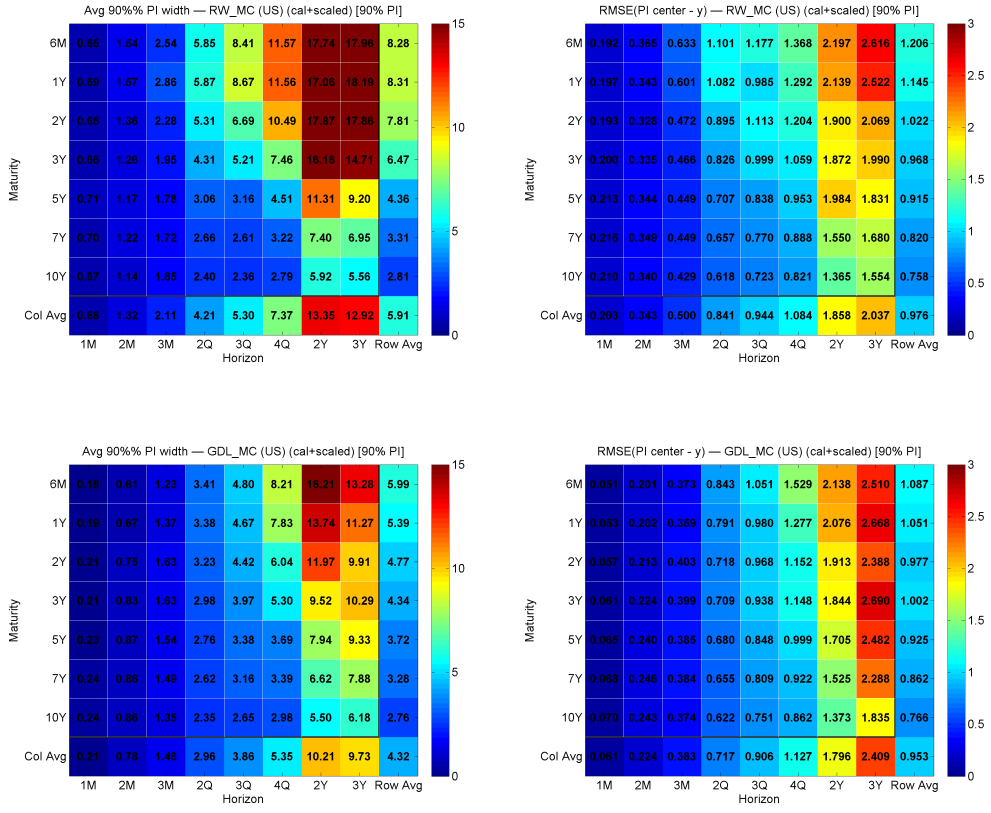


Figure D.43: Monte Carlo results for FSSM and FSSM-US and country U.S.: Average PI Widths and their RMSE at 90%

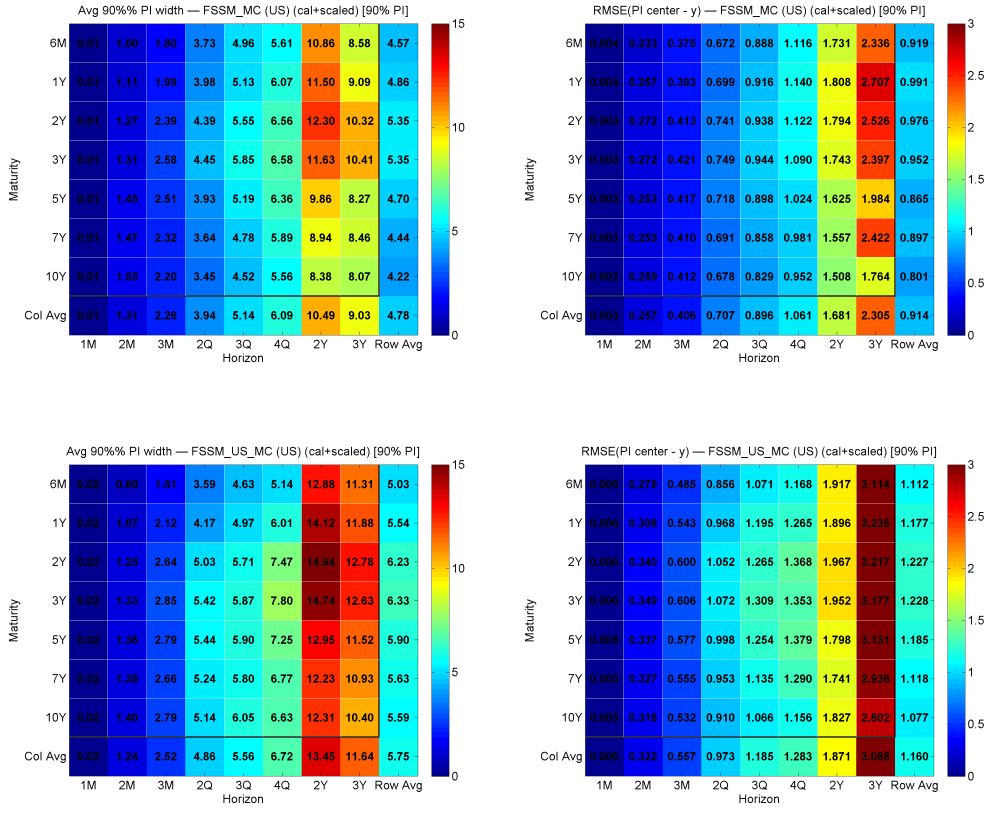


Figure D.44: Monte Carlo results for RW and GDL and country Germany: Average PI Widths and their RMSE at 90%

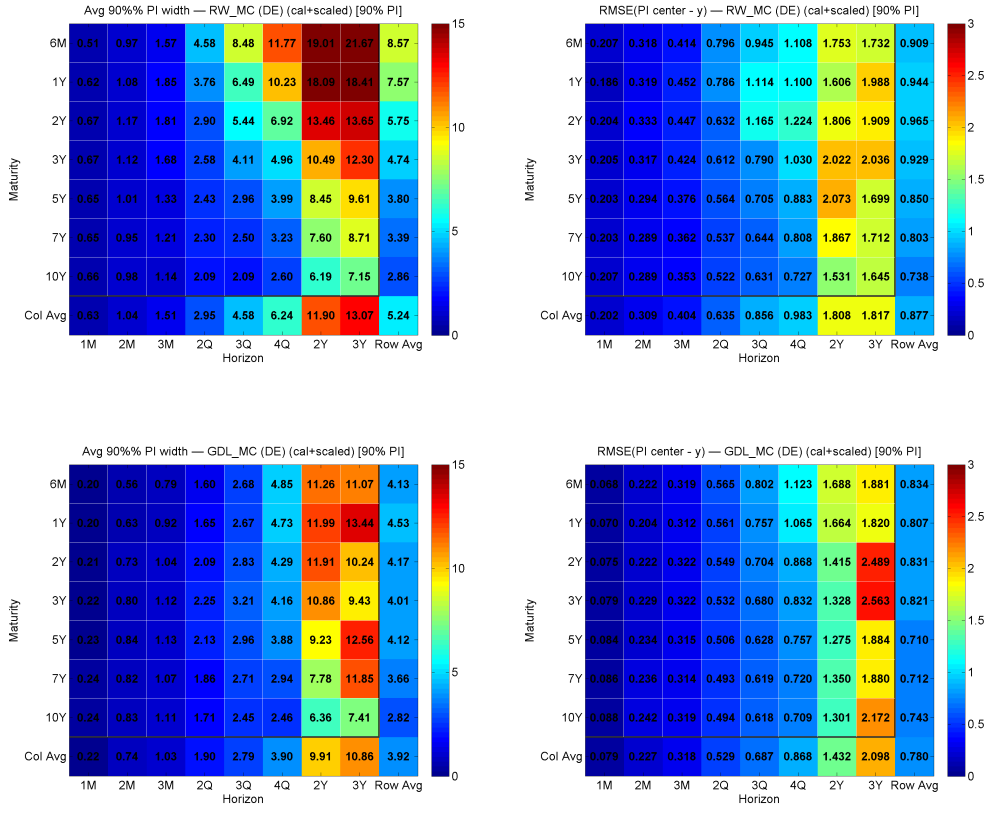


Figure D.45: Monte Carlo results for FSSM and FSSM-US and country Germany: Average PI Widths and their RMSE at 90%

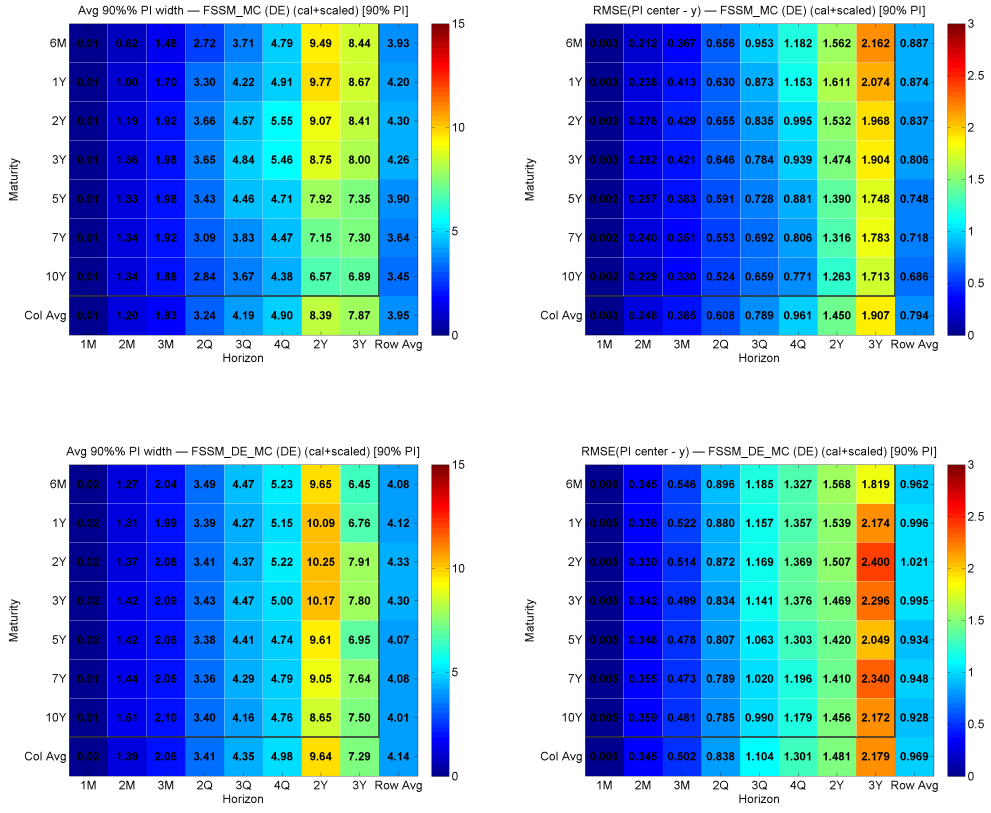


Figure D.46: Monte Carlo results for RW and GDL and country U.S.: Average PI Widths and their RMSE at 99%

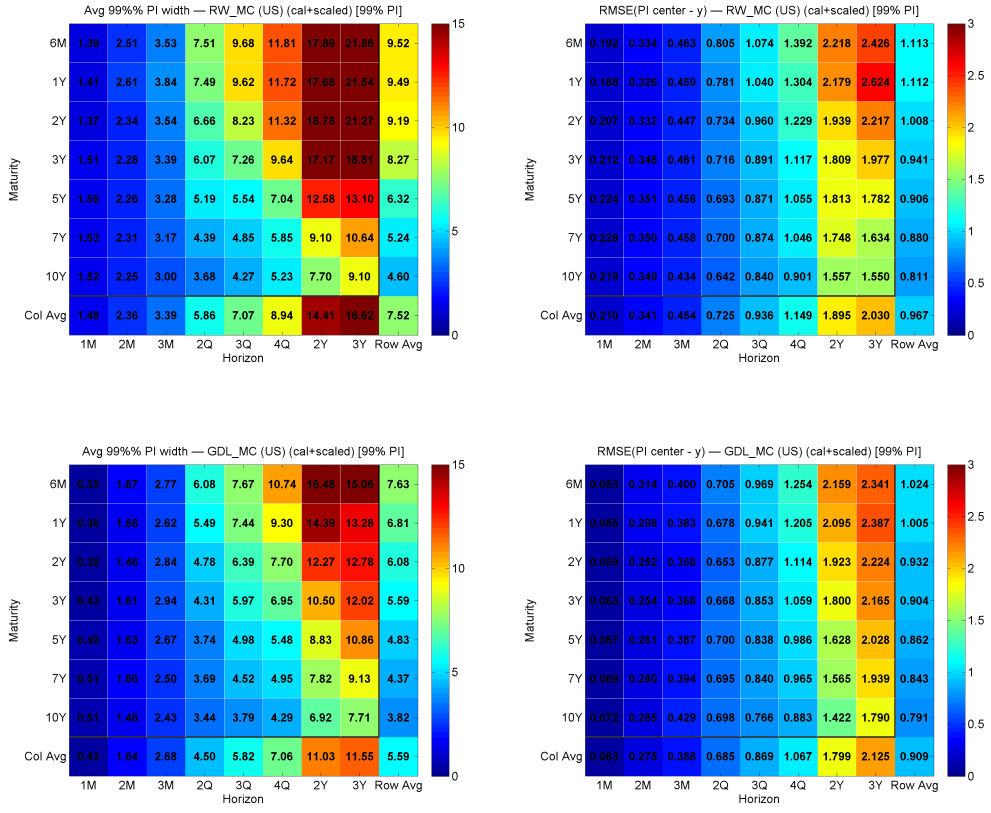


Figure D.47: Monte Carlo results for FSSM and FSSM-US and country U.S.: Average PI Widths and their RMSE at 99%

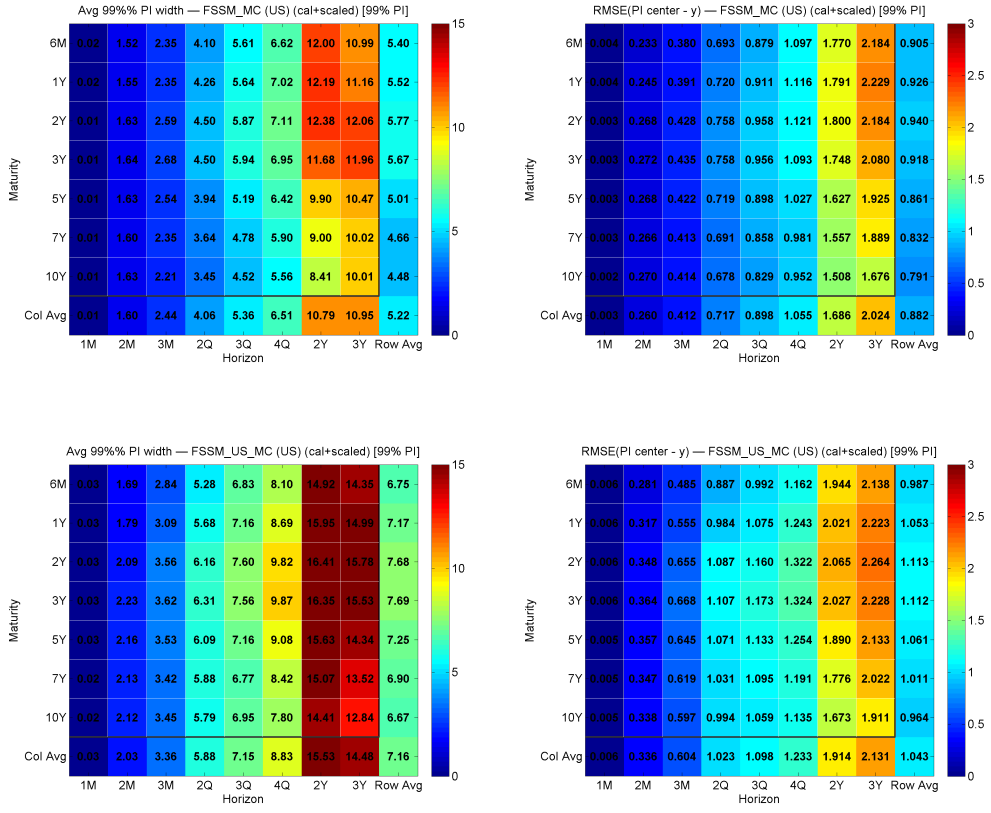


Figure D.48: Monte Carlo results for RW and GDL and country Germany: Average PI Widths and their RMSE at 99%

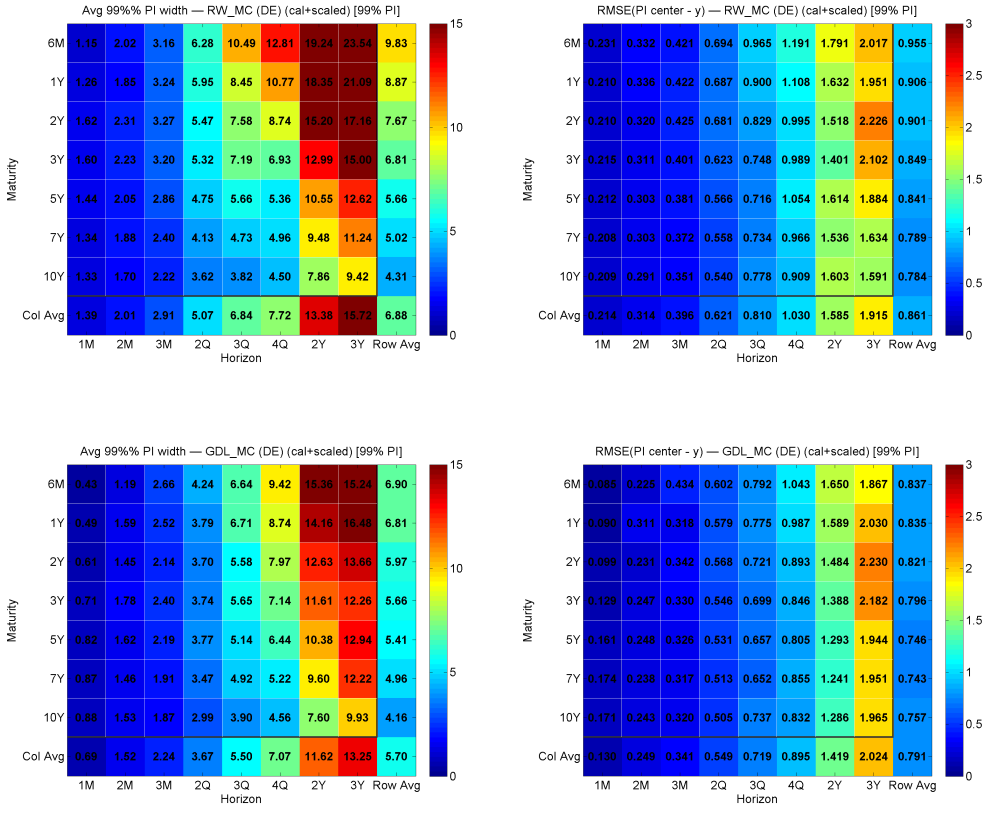
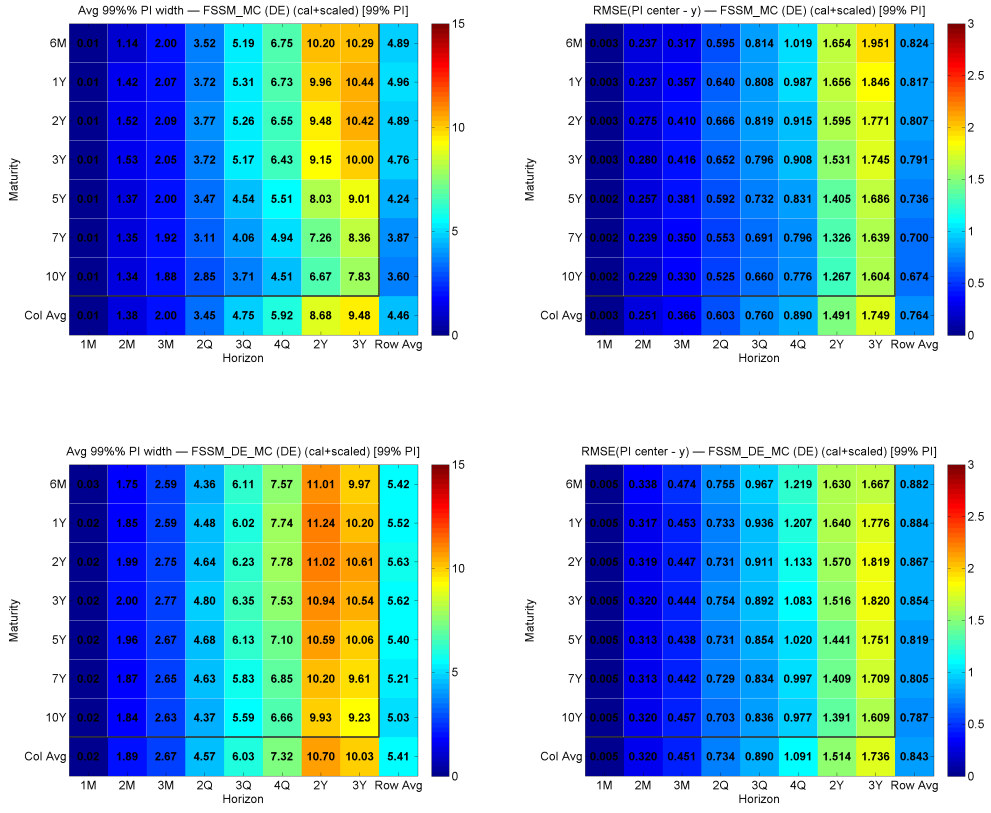


Figure D.49: Monte Carlo results for FSSM and FSSM-US and country Germany: Average PI Widths and their RMSE at 99%



References

Andersen, T. G., & Lund, J. (1997). *Stochastic volatility and mean drift in the short rate diffusion: sources of steepness, level and curvature in the yield curve*. Technical Report Working Paper, Northwestern University.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, (pp. 821–856).

Bai, J. (1994). Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis*, *15*, 453–472.

Bai, J. (1997). Estimation of a change point in multiple regression models. *The Review of Economics and Statistics*, *79*, 551–563.

Bai, J., Lumsdaine, R. L., & Stock, J. H. (1998). Testing for and dating common breaks in multivariate time series. *The Review of Economic Studies*, *65*, 395–432.

Bai, J., & Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, (pp. 47–78).

Bai, J. et al. (2000). *Vector autoregressive models with structural changes in regression coefficients and in variance-covariance matrices*. Technical Report China Economics and Management Academy, Central University of Finance and Economics.

Balduzzi, P., Das, S. R., Foresi, S., & Sundaram, R. K. (1996). A simple approach to three-factor affine term structure models. *The Journal of Fixed Income*, *6*, 43–53.

Belke, A., & Gros, D. (2005). Asymmetries in transatlantic monetary policy-making: Does the ECB follow the Fed? *JCMS: Journal of Common Market Studies*, *43*, 921–946.

Bliss, R. R. (1997a). Movements in the term structure of interest rates. *Economic Review-Federal Reserve Bank of Atlanta*, *82*, 16.

Bliss, R. R. (1997b). Testing term structure estimation methods. *Advances in Futures and Options Research 9*, (pp. 97–231).

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Brandt, M. W., & Yaron, A. (2003). *Time-consistent no-arbitrage models of the term structure*. Technical Report National Bureau of Economic Research.

Chen, L. (1996). *Stochastic mean and stochastic volatility: a three-factor model of the term structure of interest rates and its applications in derivatives pricing and risk management*. Blackwell publishers.

Commandeur, J. J., & Koopman, S. J. (2007). *An introduction to state space time series analysis*. Oxford University Press.

Commandeur, J. J., Koopman, S. J., Ooms, M. et al. (2011). Statistical software for state space methods, .

Cox, J. C., Ingersoll Jr, J. E., & Ross, S. A. (1977). A theory of the term structure of interest rates. *Econometrica*, *53*, 385–407.

Dai, Q., & Singleton, K. J. (2000). Specification analysis of affine term structure models. *The Journal of Finance*, *55*, 1943–1978.

De Jong, F., & Santa-Clara, P. (1999). The dynamics of the forward interest rate curve: A formulation with state variables. *Journal of Financial and Quantitative Analysis*, *34*, 131–157.

De Jong, P., & Penzer, J. (1998). Diagnosing shocks in time series. *Journal of the American Statistical Association*, *93*, 796–806.

De La Dehesa, G. (2013). Non-standard and unconventional monetary policy measures. *Non-Standard Monetary Policy Measures-An Update, European Parliament Directorate General for International Policies Policy Department*, (pp. 43–54).

Diebold, F. X., & Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, *130*, 337–364.

Diebold, F. X., Li, C., & Yue, V. Z. (2008). Global yield curve dynamics and interactions: a dynamic Nelson–Siegel approach. *Journal of Econometrics*, *146*, 351–363.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*, 253–263.

Diebold, F. X., & Rudebusch, G. D. (2013). *Yield curve modeling and forecasting: the dynamic Nelson-Siegel approach*. Princeton University Press.

Diebold, F. X., Rudebusch, G. D., & Aruoba, S. B. (2006). The macroeconomy and the yield curve: a dynamic latent factor approach. *Journal of Econometrics*, *131*, 309–338.

Duffee, G. R. (2002). Term premia and interest rate forecasts in affine models. *The Journal of Finance*, *57*, 405–443.

Duffie, D., & Kan, R. (1996). A yield-factor model of interest rates. *Mathematical Finance*, *6*, 379–406.

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods* volume 38. OUP Oxford.

ECB (2011a). The European Central Bank, the Eurosystem, the European System of Central Banks. *European Central Bank, Eurosystem*, .

ECB (2011b). The implementation of monetary policy in the euro area: general documentation on Eurosystem monetary policy instruments and procedures. *European Central Bank*, .

ECB (2011c). The monetary policy of the ECB. *European Central Bank, Eurosystem*, .

Fed (2018). The Federal Reserve System: Purposes & Functions. *Federal Reserve System Publication*, .

Hamilton, J. D. (1994). *Time series analysis* volume 2. Princeton University Press Princeton.

Hansen, P. R. (2003). Structural changes in the cointegrated vector autoregressive model. *Journal of Econometrics*, *114*, 261–295.

Hanspeter, S. K. (2004). The European Central Bank–history, role and functions. *European Central Bank*, .

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge university press.

Harvey, A. C. (1993). Time series models, .

Harvey, A. C., & Koopman, S. J. (1992). Diagnostic checking of unobserved-components time series models. *Journal of Business & Economic Statistics*, *10*, 377–389.

Heath, D., Jarrow, R., & Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica*, *60*, 77–105.

Hull, J., & White, A. (1990). Pricing interest-rate-derivative securities. *The Review of Financial Studies*, *3*, 573–592.

Jong, F. d. (2000). Time series and cross-section information in affine term-structure models. *Journal of Business & Economic Statistics*, *18*, 300–314.

Jotikasthira, C., Le, A., & Lundblad, C. (2015). Why do term structures in different currencies co-move? *Journal of Financial Economics*, *115*, 58–83.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, *82*, 35–45.

Litterman, R. B., & Scheinkman, J. (1991). Common factors affecting bond returns. *The Journal of Fixed Income*, *1*, 54–61.

Litzenberger, R., Squassi, G., & Weir, N. (1995). *Spline models of the term structure of interest rates and their applications*. Technical Report Working Paper, Goldman, Sachs and Company.

Mittnik, S. (1989). Multivariate time series analysis with state space models. *Computers & Mathematics with Applications*, *17*, 1189–1201.

Mittnik, S. (1990). Forecasting with balanced state space representations of multivariate distributed lag models. *Journal of Forecasting*, *9*, 207–218.

Nelson, C. R., & Siegel, A. F. (1987). Parsimonious modeling of yield curves. *Journal of Business*, (pp. 473–489).

Penzer, J. (2007). State space models for time series with patches of unusual observations. *Journal of Time Series Analysis*, *28*, 629–645.

Perron, P., Zhou, J. et al. (2008). Testing jointly for structural changes in the error variance and coefficients of a linear regression model. *Unpublished Manuscript, Department of Economics, Boston University*, .

Qu, Z., & Perron, P. (2007). Estimating and testing structural changes in multivariate regressions. *Econometrica*, *75*, 459–502.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American statistical Association*, *55*, 324–330.

Rey, H. (2016). International channels of transmission of monetary policy and the mundellian trilemma. *IMF Economic Review*, *64*, 6–35.

Ribarits, T., & Hanzon, B. (2014a). The state-space error correction model: Definition, estimation and model selection, .

Ribarits, T., & Hanzon, B. (2014b). The state-space error correction model: Simulations and applications, .

Rodriguez, C., Carrasco, C. A. et al. (2014). *ECB Policy Responses between 2007 and 2014: a chronological analysis and a money quantity assessment of their effects*. Technical Report.

Stock, J. H., & Watson, M. W. (2005). Understanding changes in international business cycle dynamics. *Journal of the European Economic Association*, *3*, 968–1006.

Toda, H. Y., & Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of econometrics*, *66*, 225–250.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics*, *5*, 177–188.

Verhelst, S. (2014). All monetary policy has become "unconventional". egmont commentary, 4 june 2014, .

Wyplosz, C. (2013). Non-standard monetary policy measures-an update. *Non-Standard Monetary Policy Measures-An Update*, (p. 7).