Understanding Growth-at-Risk: A Markov-Switching Approach*

Dario Caldara† Danilo Cascaldi-Garcia‡

Pablo Cuba-Borda§ Francesca Loria¶

March 26, 2026

Abstract

We develop a Markov-switching VAR framework for measuring and understanding Growth-at-Risk (GaR) — the left tail of the predictive distribution of economic activity. We show analytically that time-varying skewness, a necessary condition for meaningful GaR dynamics, arises naturally in Markov-switching models through three channels linking regime uncertainty to tail risk, and that endogenous transition probabilities amplify downside risk through an indirect channel absent under fixed transition probabilities. For the empirically relevant multivariate case, we develop and compare four computational methods that deliver accurate predictive moments at a fraction of the cost of brute-force simulation, and we show how the framework supports structural analysis including generalized impulse response functions for the full predictive distribution.

Keywords: Growth-at-Risk, Markov-switching, time-varying skewness, endogenous transition probabilities, predictive distribution, downside risk.

JEL codes: C32, E32, E37.

*We are grateful to Gianni Amisano, Francesco Bianchi, Richard Clarida, Domenico Giannone, Jim Hamilton, Ivana Komunjer, Junior Maih, Andrea De Michelis, Michele Modugno, Manuel Gonzalez-Astudillo, Paul Lengermann, Matthias Paustian, Giorgio Primiceri, and Andrea Raffo for their inputs to earlier versions of this work. We also thank our discussants Matteo Mogliani and Catherine Doz. The views expressed in this paper are solely those of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System.

†Federal Reserve Board, International Finance Division. dario.caldara@frb.gov.

‡Federal Reserve Board, International Finance Division. danilo.cascaldi-garcia@frb.gov.

§Federal Reserve Board, International Finance Division. pablo.a.cubaborda@frb.gov.

¶Federal Reserve Board, Monetary Affairs Division. francesca.loria@frb.gov.

Contents

1 Introduction **1**

2 Anatomy of Risk in Markov-Switching Models **4**

2.1 A Simple Markov-Switching Model 4

2.2 The Predictive Density 5

2.3 Defining Growth-at-Risk 6

2.4 Conditional Moments 8

2.5 Multi-Step Forecasts 11

2.6 An Illustration 13

3 Computing Growth-at-Risk **19**

3.1 Simulation-Based Approaches 20

3.2 An Analytical Shortcut: Probit Transitions 22

3.3 Comparison of Methods 23

3.4 Impulse Responses 23

4 Quantifying Downside Risk **27**

4.1 A Look at the Data 27

4.2 Calibration of Endogenous Transition Probabilities 28

4.3 A Bivariate Example 29

4.4 Macroeconomic Asymmetries 31

4.5 Build-up of Risk: 2007–2008 31

5 The Tightening of Lending Conditions **36**

5.1 Data and Identification 36

5.2 Impulse Responses 37
5.3 Downside Risk During the GFC 39

6 Conclusion **41**

A Proofs and Derivations for Section 2 **46**
A.1 Derivation of Path Mean and Variance 46
A.2 Proof of Proposition 1 47
A.3 Derivation of Multi-Step Conditional Moments 48

B Proofs and Derivations for Section 3 **50**
B.1 Cross-Term Integrals under Endogenous Transitions 50
B.2 Quasi-Monte Carlo: Sobol Construction and Convergence 51
B.3 Koopman Operator: Convergence and Dictionary Selection 52
B.4 The Probit-Gaussian Identity 53
B.5 Closed-Form Moments for MS-AR under Probit Transitions 54
B.6 Multivariate Normal CDF Reduction for Probit Path Weights 55
B.7 EP State Update under Probit Transitions 56
B.8 Complexity of the EP Filter 57

C MS-VAR Example with Probit Transition **57**
C.1 Model Design 57
C.2 Numerical Results 58
C.3 Forecast Density Comparison 60

D Quantile IRF Methods: Technical Details **62**
D.1 EP Forward-Recursion 62
D.2 Paired Monte Carlo Simulation 63

D.3 Validation 63

1 Introduction

Adrian, Boyarchenko and Giannone (2019) documented that the 10th percentile of the one-year-ahead predictive distribution of US real GDP growth is strongly time-varying and that its distance from the median — a measure of left-tail risk — widens sharply during periods of financial stress. This finding has spurred a literature on Growth-at-Risk (GaR), broadly defined as the outcomes associated with the left tail of the forecasted growth distribution. A central insight of that literature is that time-varying *skewness* is a necessary condition for meaningful GaR dynamics. Without it, all quantiles of the predictive distribution move in lockstep, the balance of risks remains symmetric, and the concept of GaR reduces to a simple shift in the conditional mean. Even if the 10th percentile varies over time, as long as other quantiles vary equally, the predictive distribution remains symmetric and no genuine tail risk emerges.

This paper asks what class of models can generate such time-varying skewness and, having identified it, how to compute and interpret the resulting predictive distributions in empirically relevant settings. We argue that Markov-switching models are the natural answer. Hamilton (1989) introduced the foundational framework and, as Adrian et al. (2025) emphasize, even a model with exogenous transition probabilities is capable of generating Growth-at-Risk. Timmermann (2000) derived closed-form expressions for the unconditional moments of Markov-switching models, showing that simple regime-switching specifications generate rich patterns of asymmetry and persistence. Bianchi (2016) and Kole and van Dijk (2023) subsequently derived conditional moment formulas for models with exogenous transition probabilities, under strong-persistence and weak-persistence assumptions respectively. We build on and extend these contributions in three directions.

First, we provide a unified analytical characterization of the three channels through which the predictive mixture distribution determines GaR. The predictive density at any horizon is

a mixture of Gaussians; its quantiles are driven by (i) the level of path-specific conditional means, (ii) the dispersion of those means and within-path variances, and (iii) the asymmetric concentration of mixture weights on adverse paths. Channels (i) and (ii) are available to linear models; channel (iii) is not. We show that time-varying skewness — a prerequisite for asymmetric tail risk — arises from channel (iii) through two mechanisms: a direct channel, whereby regime-specific slope differences widen as conditions deteriorate, and an indirect channel that operates only under endogenous transition probabilities. Harding and Pagan (2002) argued that a linear model can replicate standard business cycle characteristics, so there is no need to resort to Markov-switching. But Harding and Pagan did not target skewness, and our results show that replicating time-varying skewness requires the nonlinearity that regime-switching provides.

Second, we enrich the standard Markov-switching framework with endogenous transition probabilities that depend on current macroeconomic and financial conditions. Building on Diebold and Weinbach (1994) and Filardo (1994), who introduced time-varying and indicator-dependent transition probabilities respectively, our specification allows financial stress to steer the economy toward higher downside risk through two complementary mechanisms. Increased financial stress — such as unexpectedly tighter bank lending conditions — directly reduces the expected level of output in both regimes, lowering the mixture mean. At the same time, it raises the probability of transitioning into, and remaining in, the recession regime, assigning greater weight to the recessionary scenario and shifting the predictive distribution leftward while increasing its variance and left skewness. This structural interpretation of how financial conditions shape the conditional growth distribution connects our work to the evidence in Gourio (2012), Bocola (2016), Giglio, Kelly and Pruitt (2016), Adrian et al. (2022), and Fernández-Villaverde, Hurtado and Nuno (2023), but unlike the reduced-form models in that literature our Markov-switching setup allows us to identify the structural sources of tail risk. Most of the efforts to estimate macroeconomic risk have been undertaken

within reduced-form models such as quantile regressions (Adrian, Boyarchenko and Giannone, 2019; Plagborg-Møller et al., 2020), time-varying models for conditional volatility (Caldara, Scotti and Zhong, 2021), or for conditional skewness (Salgado, Guvenen and Bloom, 2019; Delle Monache, De Polis and Petrella, 2024; De Polis, Melosi and Petrella, 2024). Some structural attempts exist (Montes-Galdón and Ortega, 2022; Chavleishvili and Manganelli, 2023), but these are fairly specific and cannot be easily mapped into the standard structural VAR representations routinely used in macroeconometrics to study the effects of shocks.

Third, we develop tractable methods for computing multi-step predictive distributions in multivariate Markov-switching models with endogenous transitions, where closed-form expressions are unavailable because path weights depend on future unobserved realizations of the endogenous variables. We propose and compare four methods: standard Monte Carlo, Quasi-Monte Carlo with Sobol sequences, a Koopman operator approach, and an analytical EP filter for probit transition functions. Our framework supports a full suite of structural exercises, including generalized impulse response functions (GIRFs) for the entire predictive distribution in the spirit of Gallant, Rossi and Tauchen (1993) and Koop, Pesaran and Potter (1996). Earlier work by Rubio-Ramírez, Waggoner and Zha (2005), Sims and Zha (2006), Karamé (2010), and Karamé (2015) developed GIRFs for Markov-switching VARs under exogenous transitions, with progressively relaxed assumptions on regime knowledge at the forecast origin. We extend this to the endogenous case and to distributional rather than mean responses, connecting also to the joint distribution approach of Adrian, Boyarchenko and Giannone (2021) and Karlsson, Mazur and Nguyen (2023). DFM models with regime switches have been implemented in the literature (Camacho, Perez-Quiros and Poncela, 2018; Doz, Ferrara and Pionnier, 2020) but the derivation of macroeconomic risk and structural analysis in such an environment has been left unexplored.

The remainder of the paper is organized as follows. Section 2 establishes the analytical connection between Markov-switching models and Growth-at-Risk in the univariate case,

deriving the predictive density, conditional moments, and sources of time-varying skewness. Section 3 extends the framework to the MS-VAR, characterizes the analytical obstruction under endogenous transitions, and develops four computational methods together with guidance on calibrating the transition probabilities. Section 4 applies the bivariate model to U.S. data, documenting time-varying asymmetries in the predictive distributions of industrial production and financial conditions and tracing the build-up of fragility during the Global Financial Crisis. Section 5 extends the framework to a six-variable structural MS-VAR and studies the macroeconomic transmission of a tightening of bank lending conditions. Section 6 concludes.

2 Anatomy of Risk in Markov-Switching Models

In this section we establish the analytical connection between Markov-switching models and Growth-at-Risk. We set up the model, derive the predictive density and conditional moments, define GaR and its three channels, identify the sources of time-varying skewness, and extend the framework to multi-step forecasts.

2.1 A Simple Markov-Switching Model

Let $s_t \in \{1, 2\}$ denote the latent regime at time t, with state 1 representing economic expansions and state 2 recessions. The observed variable follows a Markov-switching autoregression with one lag—a MS-AR(1):

$$y_t = \mu_{s_t} + \phi_{s_t}\, y_{t-1} + \sigma_{s_t}\, \eta_t, \qquad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \tag{1}$$

where μ_{s_t}, ϕ_{s_t}, and σ_{s_t} are the regime-specific intercept, autoregressive slope, and innovation standard deviation. The innovation η_t captures idiosyncratic *shock uncertainty* within each regime, independently of the regime sequence.

The latent regime evolves as a Markov chain with transition probabilities

$$p_{jk}(y_t) \equiv \Pr(s_{t+1} = k \mid s_t = j,\, y_t), \quad j, k \in \{1, 2\}, \quad \sum_{k=1}^{2} p_{jk}(y_t) = 1. \tag{2}$$

Transitions are *exogenous* when $p_{jk}(y_t) = p_{jk}$ are fixed constants, and *endogenous* when they depend on the current state y_t. A natural specification, discussed further in Section 2.4, is the logistic form, which ensures that deteriorating conditions raise the probability of entering the recession regime and prolong recovery from it.

Let $\mathcal{F}_t = \{y_t, y_{t-1}, \ldots\}$ be the information set at time t. The filtered regime probability $\pi_{j,t} \equiv \Pr(s_t = j \mid \mathcal{F}_t)$ is a sufficient statistic for forecasting: all objects of interest depend on $\mathcal{F}_t$ only through the pair $(y_t, \pi_{j,t})$. The one-step-ahead predicted regime probability is

$$\pi_{k,t+1|t} \equiv \Pr(s_{t+1} = k \mid \mathcal{F}_t) = \sum_{j=1}^{2} p_{jk}(y_t)\, \pi_{j,t}. \tag{3}$$

2.2 The Predictive Density

The object of interest is the conditional predictive density of y_{t+h} given $\mathcal{F}_t$, for $h \geq 1$. At $h = 1$, marginalizing over the next-period regime yields a finite Gaussian mixture:

$$f(y_{t+1} \mid \mathcal{F}_t) = \sum_{k=1}^{2} \pi_{k,t+1|t} \cdot \mathcal{N}\left(\mu_k + \phi_k\, y_t,\ \sigma_k^2\right). \tag{4}$$

Each component is Gaussian with regime-specific mean and variance; the mixture can be non-Gaussian depending on conditions we characterize below. Under endogenous transitions, the mixture weights $\pi_{k,t+1|t}$ respond to y_t through (3), making the predictive distribution state-dependent.

At longer horizons $h > 1$, the predictive density preserves this mixture structure but integrates over all possible regime *paths* between t and $t + h$. A regime path is a sequence

$\boldsymbol{\kappa} = (k_1, \ldots, k_h) \in \{1, 2\}^h$; there are 2^h such paths. Conditional on a path $\boldsymbol{\kappa}$ and $\mathcal{F}_t$, the model iterates as a sequence of linear Gaussian steps, so $y_{t+h} \mid \boldsymbol{\kappa}, \mathcal{F}_t$ is Gaussian with mean $m_{\boldsymbol{\kappa}}$ and variance $v_{\boldsymbol{\kappa}}$ satisfying the recursion

$$m_{\boldsymbol{\kappa}} = \Big(\prod_{\ell=1}^{h} \phi_{k_\ell}\Big) y_t + \sum_{j=1}^{h} \Big(\prod_{\ell=j+1}^{h} \phi_{k_\ell}\Big) \mu_{k_j}, \tag{5}$$

$$v_{\boldsymbol{\kappa}} = \sum_{j=1}^{h} \Big(\prod_{\ell=j+1}^{h} \phi_{k_\ell}^2\Big) \sigma_{k_j}^2, \tag{6}$$

with empty products equal to one (see Appendix A.1 for a proof by induction). Marginalizing over paths gives the h-step-ahead predictive density as a mixture of 2^h Gaussians:

$$f(y_{t+h} \mid \mathcal{F}_t) = \sum_{\boldsymbol{\kappa} \in \{1,2\}^h} w_{\boldsymbol{\kappa}}\, \mathcal{N}(m_{\boldsymbol{\kappa}},\, v_{\boldsymbol{\kappa}})\,, \tag{7}$$

where $w_{\boldsymbol{\kappa}} = \Pr(s_{t+1} = k_1, \ldots, s_{t+h} = k_h \mid \mathcal{F}_t)$ are the *path weights*. Setting $h = 1$ recovers (4) with $w_k = \pi_{k,t+1|t}$.

2.3 Defining Growth-at-Risk

Growth-at-Risk (GaR) at horizon h and confidence level $\tau \in (0, 1)$ is the τ-quantile of the predictive distribution:

$$\mathrm{GaR}_{t,h}(\tau) \equiv \inf\Big\{q : F_{t,h}(q) \geq \tau\Big\}, \tag{8}$$

where $F_{t,h}(q) = \Pr(y_{t+h} \leq q \mid \mathcal{F}_t)$ is the h-step-ahead predictive CDF. Conventional choices are $\tau = 0.05$ or $\tau = 0.10$. Since the predictive density is a mixture of 2^h Gaussians,

$$F_{t,h}(q) = \sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}}\, \Phi\left(\frac{q - m_{\boldsymbol{\kappa}}}{\sqrt{v_{\boldsymbol{\kappa}}}}\right), \tag{9}$$

where Φ is the standard normal CDF. The GaR has no closed form in q, but its determinants are transparent from the mixture structure (9). Three forces push $\text{GaR}_{t,h}(\tau)$ downward, worsening downside risk:

1. **Lower path means** $m_{\boldsymbol{\kappa}}$ shift the mixture (9) leftward, reducing all quantiles equally. This is the only channel available to a linear model.

2. **Greater dispersion in** $m_{\boldsymbol{\kappa}}$ **or larger** $v_{\boldsymbol{\kappa}}$ fattens both tails of (9) symmetrically, lowering the left quantile for any $\tau < 0.5$ but raising the right quantile by the same amount.

3. **Asymmetric concentration of** $w_{\boldsymbol{\kappa}}$ on paths with low $m_{\boldsymbol{\kappa}}$ skews (9) to the left at fixed mean and dispersion, lowering the left quantile *without* a corresponding improvement in the right tail.

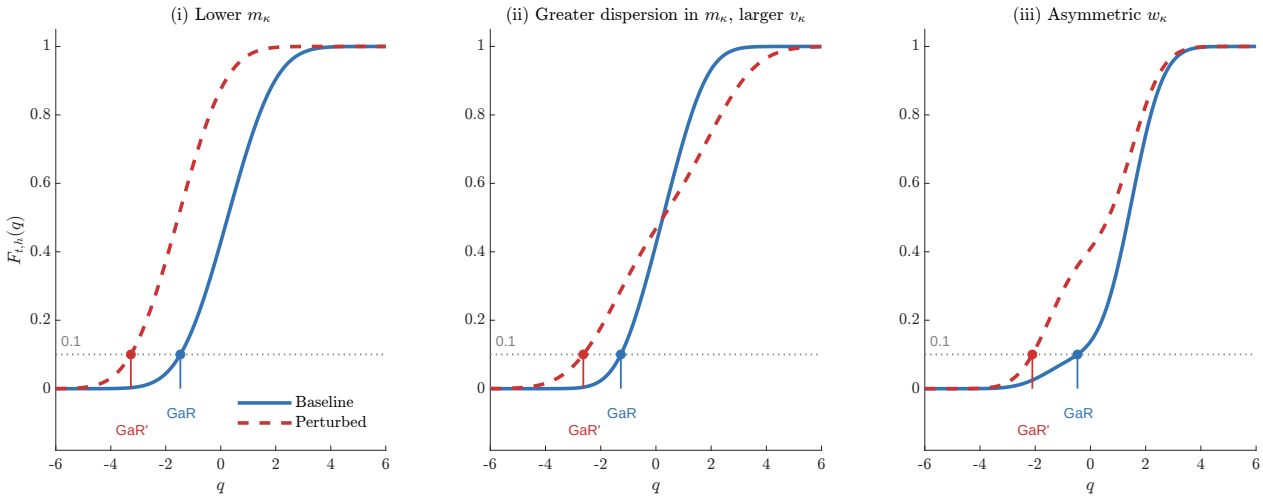


Figure 1: Three channels through which the predictive mixture CDF determines Growth-at-Risk. Each panel plots the baseline CDF (solid blue) and a perturbed CDF (dashed red), with the implied $\text{GaR}(\tau = 0.10)$ marked by vertical lines. *Panel (i):* all path means $m_{\boldsymbol{\kappa}}$ shift down by 1.8 units; the CDF translates leftward rigidly. *Panel (ii):* path means spread further apart and path variances increase, with the mixture mean re-centered; the CDF widens symmetrically. *Panel (iii):* path weights shift from $(w_1, w_2) = (0.1, 0.9)$ to $(0.4, 0.6)$, concentrating mass on the low-mean path ($m_1 = -1.5$, $m_2 = 1.5$); the left tail thickens while the right tail barely moves, generating asymmetric deterioration in downside risk. Channels (i) and (ii) are available to linear models; channel (iii) is not.

Figure 1 illustrates these three channels. In each panel, the baseline CDF (solid blue) corresponds to a mixture with most weight on the high-mean path, and the perturbed CDF (dashed red) isolates one channel by holding all other features of the mixture fixed. Panel (i) shifts all path means $m_{\boldsymbol{\kappa}}$ downward: the CDF translates leftward rigidly and GaR worsens, but so does the upper quantile by the same amount. Panel (ii) widens the dispersion in $m_{\boldsymbol{\kappa}}$ and raises $v_{\boldsymbol{\kappa}}$, holding the mixture mean fixed: both tails fatten symmetrically and the balance of risks is again unchanged. Panel (iii) shifts the path weights toward the low-mean path, holding means and variances fixed: the left tail thickens while the right tail barely moves. Channel (iii) is absent in any model with a symmetric predictive distribution and is the distinctive contribution of the regime-switching structure. We formalize this observation in the remark below.

Remark 1 (Necessity of time-varying skewness)**.** *For any predictive distribution that is symmetric at every date and horizon, channels (i) and (ii) remain available but channel (iii) is permanently absent. The τth and $(1-\tau)$th quantiles then deviate equally from the median at all times: GaR and its upper counterpart move in lockstep, and the model generates no asymmetric tail risk. Time-varying skewness is therefore a necessary condition for meaningful GaR dynamics. Linear Gaussian models are symmetric at all horizons by construction and cannot satisfy this condition. Markov-switching models are generically asymmetric whenever regimes differ in means or variances, and they satisfy it through the mechanisms characterized in Section 2.4.*

2.4 Conditional Moments

The mixture structure (4) yields closed-form expressions for all one-step-ahead conditional moments. To highlight the sources of time variation, define the *regime mean gap*—the

deviation of regime k's conditional mean from the mixture mean:

$$\begin{aligned} d_k &= (\mu_k + \phi_k\, y_t) - \sum_{j=1}^{2} \pi_j (\mu_j + \phi_j\, y_t) \\ &= \underbrace{(\mu_k - \bar{\mu})}_{\text{intercept dispersion}} + \underbrace{(\phi_k - \bar{\phi})\, y_t}_{\text{persistence dispersion}}, \end{aligned} \tag{10}$$

where $\bar{\mu} = \sum_k \pi_k \mu_k$ and $\bar{\phi} = \sum_k \pi_k \phi_k$. The gap has two structurally distinct components. The *intercept dispersion* $\mu_k - \bar{\mu}$ creates a baseline wedge between regimes that depends on the mixture weights but not directly on the level of y_t. The *persistence dispersion* $(\phi_k - \bar{\phi})\, y_t$ grows linearly with $|y_t|$ whenever slopes differ across regimes. Both components shape the variance and skewness of the predictive distribution.

Proposition 1 (Conditional moments)**.** *Write $\pi_k \equiv \pi_{k,t+1|t}$. The one-step-ahead conditional mean, variance, and skewness of the predictive distribution (4) are:*

$$\mathbb{E}[y_{t+1} \mid \mathcal{F}_t] = \bar{\mu} + \bar{\phi}\, y_t, \tag{11}$$

$$\mathbb{V}[y_{t+1} \mid \mathcal{F}_t] = \underbrace{\sum_k \pi_k\, \sigma_k^2}_{\textit{shock uncertainty}} + \underbrace{\sum_k \pi_k\, d_k^2}_{\textit{regime uncertainty}}, \tag{12}$$

$$\text{Skew}(y_{t+1} \mid \mathcal{F}_t) = \frac{\overbrace{3 \sum_k \pi_k\, \sigma_k^2\, d_k}^{\textit{cross term}} + \overbrace{\sum_k \pi_k\, d_k^3}^{\textit{between term}}}{\mathbb{V}[y_{t+1} \mid \mathcal{F}_t]^{3/2}}. \tag{13}$$

Proof. See Appendix A.2. □

Remark 2 (Extension to the MS-VAR)**.** *Proposition 1 extends directly to the MS-VAR (19): scalar quantities are replaced by their vector counterparts, but the decomposition of forecast uncertainty into shock and regime components, and the sources of time-varying skewness, are identical.*

The conditional mean is the probability-weighted average of regime-specific predicted values $\mu_k + \phi_k y_t$, responding linearly to y_t with effective slope $\bar{\phi}$. The variance decomposes into *shock uncertainty*—the average within-regime variance, which varies over time when σ_k^2 differs across regimes and the weights π_k move—and *regime uncertainty*—the dispersion of regime-specific conditional means, which is positive whenever regimes have different conditional means and weights are interior.

The skewness numerator departs from the variance in one key respect: there is no within-regime component, because each regime is Gaussian and therefore internally symmetric. The *between term* $\sum_k \pi_k d_k^3$ captures pure location asymmetry; for two regimes it equals $\pi_1 \pi_2 (m_1 - m_2)^3 (\pi_2 - \pi_1)$, which is negative when the expansion regime dominates ($\pi_1 > \pi_2$), zero at equal weights, and positive when the recession regime dominates. The *cross term* $3 \sum_k \pi_k \sigma_k^2 d_k$ captures the interaction between within-regime variance and mean displacement; it contributes negative skewness when the high-variance regime has the lower conditional mean—the empirically relevant case in which the recession regime has both higher volatility and lower growth. This interaction has no counterpart in the variance decomposition: variance is insensitive to the sign of d_k, while skewness is not.

Both terms respond to y_t through two channels. The *direct channel* operates even under exogenous transitions: whenever slopes differ across regimes, the persistence dispersion term $(\phi_k - \bar{\phi})\, y_t$ in d_k widens as y_t falls, increasing the between and cross terms and pushing skewness more negative. The *indirect channel* requires endogenous transitions: as y_t deteriorates, the weights $\pi_{k,t+1|t}(y_t)$ shift toward the recession regime, which simultaneously changes the summation weights in every moment formula and shifts the mixture mean $\bar{\mu}(y_t) = \sum_k \pi_{k,t+1|t}(y_t)\, \mu_k$, making the intercept dispersion $\mu_k - \bar{\mu}$ in d_k state-dependent even when all structural parameters are fixed. The two channels reinforce each other: deteriorating conditions widen both the persistence and intercept dispersion components of d_k simultaneously, raising the between and cross terms together and generating a convex amplification of downside risk

that is absent under exogenous transitions.

Proposition 1 nests several results from the literature as special cases. Under exogenous transitions, $p_{jk}(y_t) = p_{jk}$ are constants and $\pi_{k,t+1|t}$ depends on $\mathcal{F}_t$ only through the filtered probabilities $\pi_{j,t}$, not directly on y_t: the indirect channel is absent and the moment expressions reduce to those of Bianchi (2016) and Kole and van Dijk (2023). Evaluating further at ergodic weights $\pi_{k,t+1|t} = \pi_k^\infty$ recovers the unconditional moments of Timmermann (2000).

Finally, we characterize the conditions under which skewness is exactly zero in Remark 3 below.

Remark 3 (Zero-skewness conditions)**.** *Skewness is exactly zero under two distinct conditions. First, if $d_k = 0$ for all k—which occurs when $\mu_1 = \mu_2$ and $\phi_1 = \phi_2$—both terms in the numerator vanish regardless of variance heterogeneity. Second, if $\pi_1 = \pi_2 = 1/2$ and $\sigma_1^2 = \sigma_2^2$: equal weights force $d_1 = -d_2$, making the between term vanish by antisymmetry, and equal variances make the cross term $\frac{3}{2}d_1(\sigma_1^2 - \sigma_2^2)$ vanish as well.*

2.5 Multi-Step Forecasts

The variance and skewness decompositions in Proposition 1 generalize directly to multi-step forecasts. Letting $d_{\boldsymbol{\kappa}} \equiv m_{\boldsymbol{\kappa}} - \bar{m}_h$ denote path mean deviations from the overall h-step predictive mean, the moments of the mixture (7) are (see derivations in Appendix A.3):

$$\bar{m}_h = \sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}}\, m_{\boldsymbol{\kappa}}, \tag{14}$$

$$\mathbb{V}[y_{t+h} \mid \mathcal{F}_t] = \underbrace{\sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}}\, v_{\boldsymbol{\kappa}}}_{\text{shock uncertainty}} + \underbrace{\sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}}\, d_{\boldsymbol{\kappa}}^2}_{\text{regime uncertainty}}, \tag{15}$$

$$\mathbb{E}\Big[(y_{t+h} - \bar{m}_h)^3 \mid \mathcal{F}_t\Big] = \sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}}\Big(3\, v_{\boldsymbol{\kappa}}\, d_{\boldsymbol{\kappa}} + d_{\boldsymbol{\kappa}}^3\Big). \tag{16}$$

The structure mirrors the one-step case: the two-regime mixture weights π_k are replaced by path weights $w_{\boldsymbol{\kappa}}$ over 2^h paths, and within-regime variances σ_k^2 are replaced by path variances $v_{\boldsymbol{\kappa}}$. Setting $h = 1$ recovers (11)–(13) exactly.

The key distinction between exogenous and endogenous transitions is whether the path weights $w_{\boldsymbol{\kappa}}$ are known at the forecast origin. Under exogenous transitions, $p_{jk}(y_t) = p_{jk}$ are fixed constants and the weight factorizes as

$$w_{\boldsymbol{\kappa}} = \pi_{k_1,t+1|t} \cdot p_{k_1 k_2} \cdots p_{k_{h-1} k_h}, \tag{17}$$

with every factor observed at t, so (14)–(16) reduce to finite weighted sums computable in closed form for any h. Under endogenous transitions the path weights depend on future, unobserved realizations of y and must be evaluated numerically. Section 3 discusses tractable approaches.

Regardless of the method used to construct the density, the decompositions (15)–(16) hold. The regime-uncertainty component is present in both variance and skewness at all horizons and does not vanish as h grows: under stationarity, path means converge to terminal-regime unconditional values $\mu_k/(1-\phi_k)$, and the regime-uncertainty term converges to the cross-sectional variance of these values weighted by ergodic probabilities π_k^∞—a strictly positive constant whenever regimes have different unconditional means. Unlike the conditional mean, which converges to its ergodic level and carries no permanent component, the variance and skewness retain a permanent regime-uncertainty contribution that does not decay with the horizon.[1]

[1]Long-horizon GaR consequently does not converge to a Gaussian quantile: the limiting predictive distribution is an ergodic mixture of Gaussians, which is generically non-Gaussian and skewed whenever the ergodic regime probabilities are asymmetric.

2.6 An Illustration

We illustrate the framework using a univariate Markov-switching autoregression (1) with two regimes and all parameters switching. The calibration loosely reflects the behavior of US industrial production during NBER expansions and recessions: the recession regime has a lower mean and higher volatility. We set $\mu_1 = 0.5$, $\mu_2 = -1.0$, $\sigma_1 = 1.0$, $\sigma_2 = 2.0$, and $\phi_1 = \phi_2 = 0.8$. Endogenous transition probabilities follow the logistic specification

$$p_{12}(y_t) = \frac{1}{1 + e^{\,\alpha_{12} + \beta_{12}\, y_t}}, \qquad p_{21}(y_t) = \frac{1}{1 + e^{\,\alpha_{21} - \beta_{21}\, y_t}}, \tag{18}$$

with $\beta_{12}, \beta_{21} > 0$, so that deteriorating conditions raise the probability of entering the recession regime and slow recovery from it. We set $\alpha_{12} = 2$, $\alpha_{21} = -2$, and $\beta_{12} = \beta_{21} = 3$. Figure 2 plots both functions. The probability of entering the recession regime, $p_{12}(y_t)$, is decreasing in y_t, while the probability of exiting it, $p_{21}(y_t)$, is increasing. With the chosen parameters, both cross the 0.5 threshold at $y_t^* = -\alpha_{12}/\beta_{12} = -2/3$, so the region of maximum state-sensitivity is the same for both transitions.

For comparison, we calibrate an exogenous transition probability model by setting p_{12} and p_{21} equal to the ergodic averages of $p_{12}(y_t)$ and $p_{21}(y_t)$ implied by the endogenous model, ensuring both specifications share the same unconditional frequency of recession. Both models are initialized at $y_t = 0$, with bad-regime probability $\pi_{2,t} = p_{12}(0)/\big(p_{12}(0) + p_{21}(0)\big) \approx 0.12$.

Time variation in the predictive distribution. Figure 3 traces the predictive distribution over time by using the observed sequence $\{y_t, \pi_{2,t}\}$ as a rolling jumpoff point. At each date t, we compute h-step-ahead predictive quantiles for both the endogenous and exogenous models and report the 10th, 50th, and 90th percentiles together with the downside asymmetry measure $(Q_{50} - Q_{10})$.

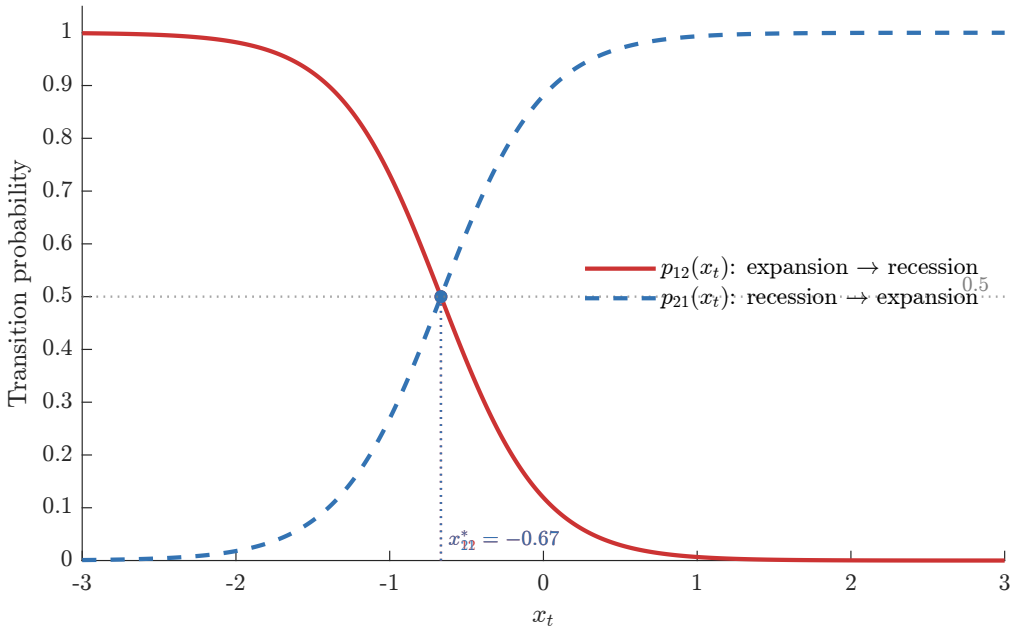


Figure 2: Logistic transition probability functions. The solid red curve shows $p_{12}(y_t)$, the probability of transitioning from expansion to recession; the dashed blue curve shows $p_{21}(y_t)$, the probability of transitioning from recession to expansion. Parameters: $\alpha_{12} = 2$, $\alpha_{21} = -2$, $\beta_{12} = \beta_{21} = 3$. Dotted vertical lines mark the switching threshold $y_t^* = -2/3$ where each probability equals 0.5.

Three features stand out. First, the left tail deteriorates markedly more than the right tail improves during bad times: when $\pi_{2,t}$ rises, the 10th percentile drops sharply while the 90th percentile barely moves, generating the negative asymmetry visible in the bottom panel. Second, the endogenous model amplifies this asymmetry relative to the exogenous model: the indirect channel concentrates probability mass on downside paths precisely when conditions deteriorate, pushing the left tail further than fixed transition probabilities would imply. Third, the asymmetry under exogenous transitions is muted and less state-dependent, consistent with the theoretical result that the indirect channel is absent when $p_{jk}(y_t) = p_{jk}$.

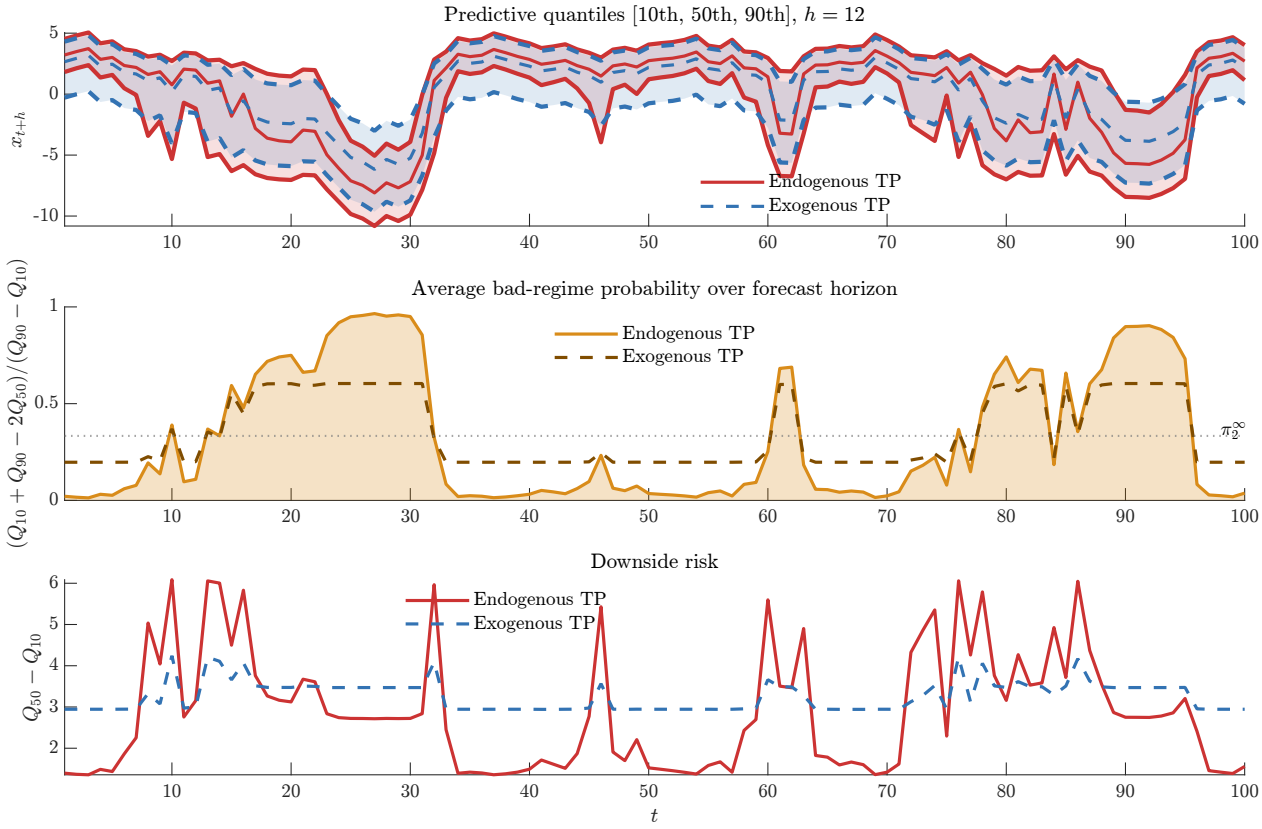


Figure 3: Time variation in the predictive distribution. *Top panel:* 10th, 50th, and 90th percentiles of the h-step-ahead predictive distribution for the endogenous (solid red) and exogenous (dashed blue) transition probability models; shaded bands cover the interquantile range. The observed series y_t is shown in gray. *Middle panel:* average bad-regime probability over the forecast horizon, $\bar{\pi}_{2,t+h|t} = h^{-1} \sum_{\ell=1}^{h} \pi_{2,t+\ell|t}$; dotted line marks the ergodic level π_2^∞. *Bottom panel:* downside risk measure $Q_{50} - Q_{10}$, the distance from the median to the 10th percentile. Higher values indicate a thicker left tail. The endogenous model generates systematically larger downside risk than the exogenous model, reflecting amplification through the indirect channel. Parameters: $\mu_1 = 0.5$, $\mu_2 = -1$, $\sigma_1 = 1$, $\sigma_2 = 2$, $\phi_1 = \phi_2 = 0.8$, $\alpha_{12} = 2$, $\alpha_{21} = -2$, $\beta_{12} = \beta_{21} = 3$.

Experiment. To illustrate how the three channels of Section 2.3 operate dynamically, we consider two scenarios: a *moderate deterioration*, defined as a one-standard-deviation adverse shock to y_t hitting at $t+1$, and a *severe deterioration*, defined as a two-standard-deviation shock of the same sign. For each scenario we track the evolution of the predictive density at horizons $h \in \{1, 2, 3\}$, reporting the conditional mean, variance, skewness, and GaR($\tau = 0.10$). The comparison between the two scenarios isolates the nonlinear amplification induced by endogenous transition probabilities: under a logistic specification, a larger shock disproportionately raises $p_{12}(y_t)$ and depresses $p_{21}(y_t)$, so the severe deterioration generates more than twice the downside risk of the moderate one. This asymmetry — absent under exogenous transitions or in any linear model — is the quantitative manifestation of channel (iii) characterized in Remark 1.

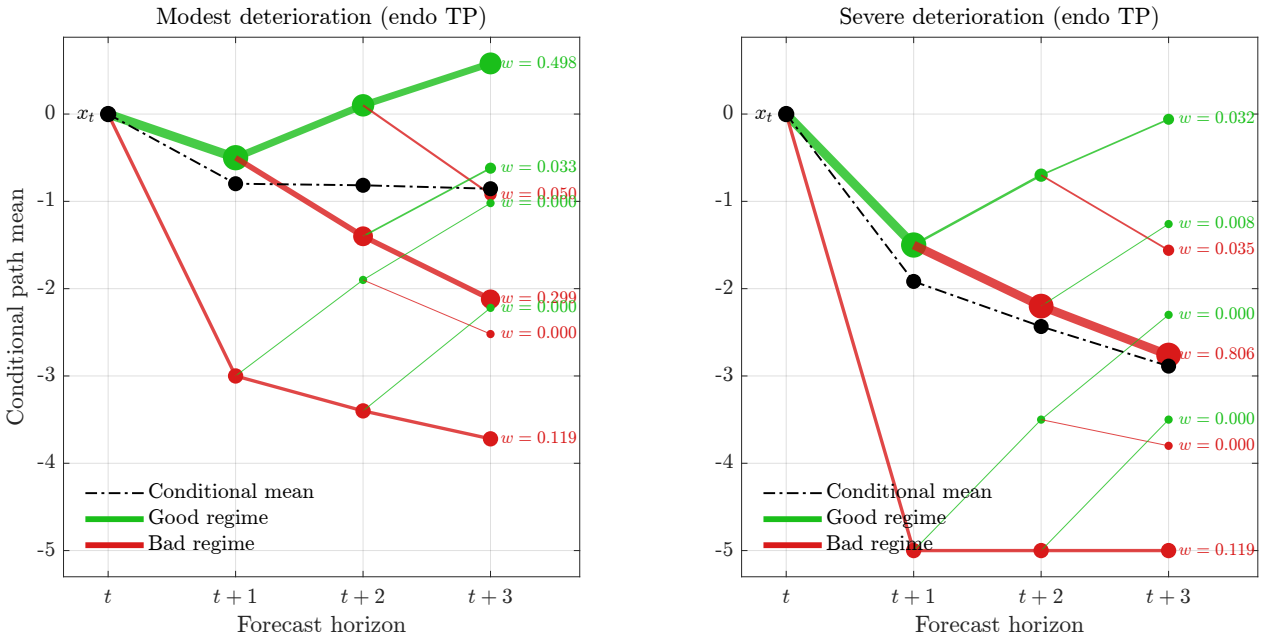


Figure 4: Bifurcation of conditional path mean.

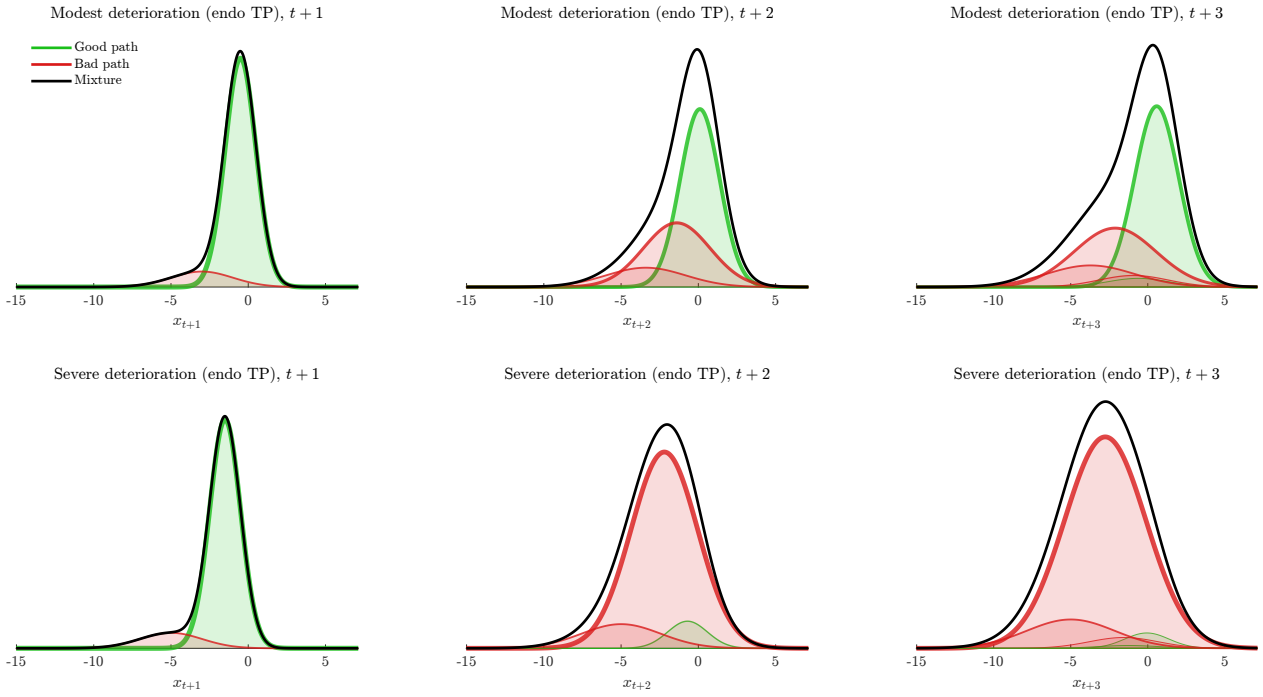
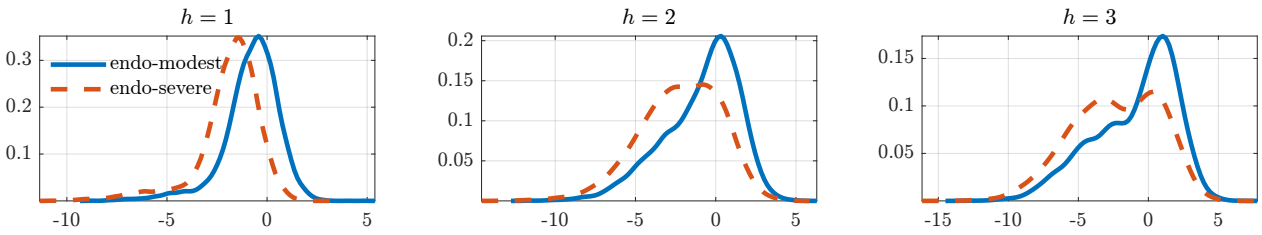
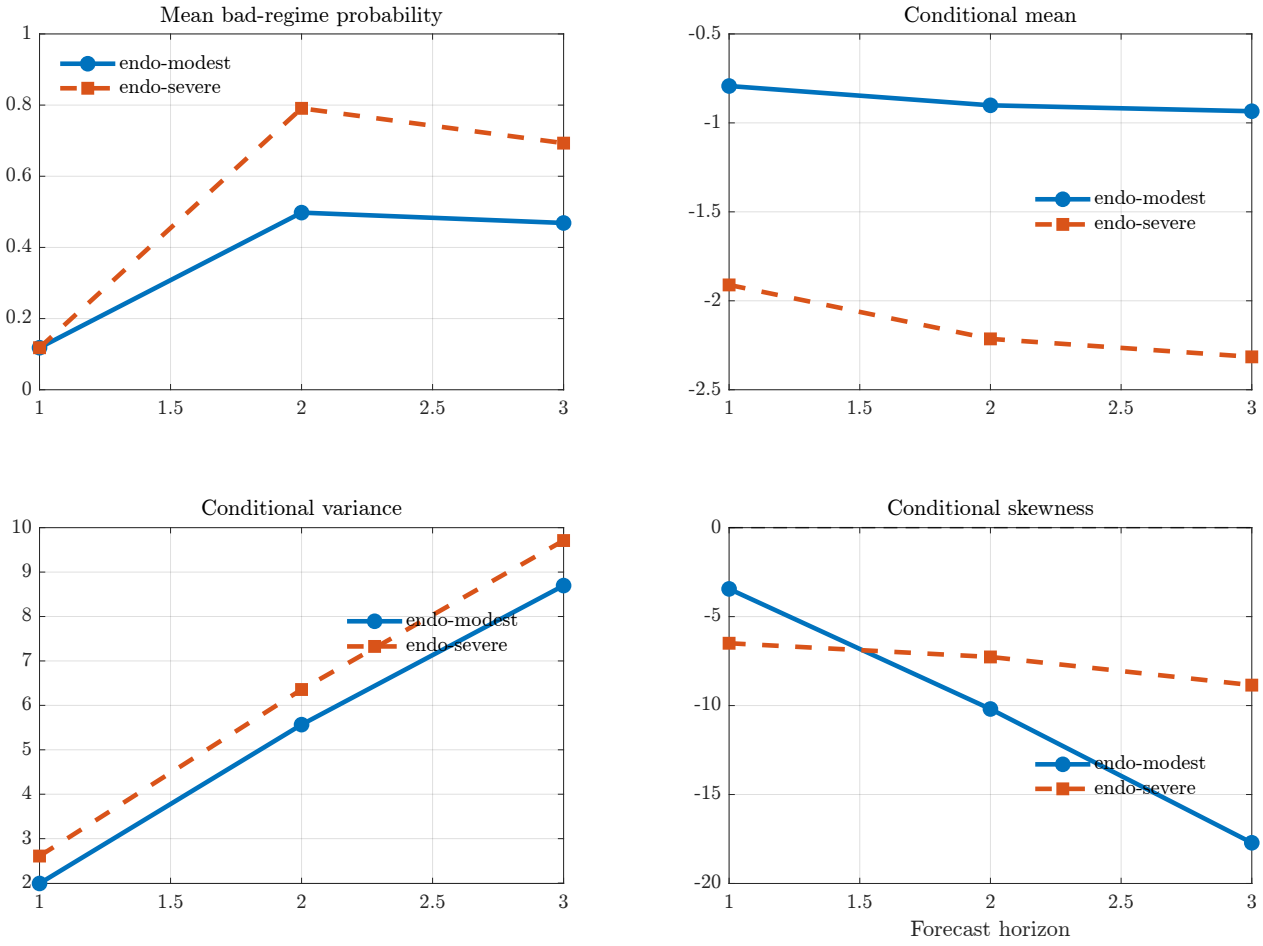


Figure 5: Mixture of densities under modest and severe deterioration.

Figure 6: Predictive densities and their conditional moments under modest (blue) vs. severe (orange) scenario.



(a) Densities



(b) Moments

3 Computing Growth-at-Risk

Section 2 established that Growth-at-Risk is determined by the path weights $w_{\boldsymbol{\kappa}}$, path means $m_{\boldsymbol{\kappa}}$, and path variances $v_{\boldsymbol{\kappa}}$ of the predictive mixture. Under exogenous transitions these objects are computable in closed form. Under endogenous transitions they are not, because the weight assigned to each regime path depends on future, unobserved realizations of the endogenous variables. This section develops tractable methods for the empirically relevant case.

In practice, the object of interest is not the univariate MS-AR(1) (1) but its multivariate generalization. Consider the structural Markov-switching VAR:

$$\mathbf{A}_0(s_t)\mathbf{y}_t = \mathbf{c}(s_t) + \mathbf{A}_1(s_t)\mathbf{y}_{t-1} + \cdots + \mathbf{A}_p(s_t)\mathbf{y}_{t-p} + \boldsymbol{\Sigma}(s_t)\,\boldsymbol{\varepsilon}_t, \tag{19}$$

where $\mathbf{y}_t$ is the $n \times 1$ vector of endogenous variables, $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ is a vector of mutually uncorrelated structural innovations, and the latent regime $s_t \in \{1, \ldots, K\}$ evolves with transition probabilities $p_{jk}(\mathbf{y}_t) \equiv p(s_{t+1} = k \mid s_t = j, \mathbf{y}_t)$. Writing $\mathbf{x}_t' = [\mathbf{1}, \mathbf{y}_{t-1}', \ldots, \mathbf{y}_{t-p}']'$ and $\mathbf{A}^+(s_t) = [\mathbf{c}(s_t), \mathbf{A}_1(s_t), \ldots, \mathbf{A}_p(s_t)]$, the model reads compactly as $\mathbf{A}_0(s_t)\mathbf{y}_t = \mathbf{A}^+(s_t)\mathbf{x}_t + \boldsymbol{\Sigma}(s_t)\boldsymbol{\varepsilon}_t$. The univariate MS-AR is a special case with $n = 1$, $p = 1$.

Under endogenous transitions, $p_{jk}(\mathbf{y}_{t+\ell})$ depends on the as-yet-unobserved realization of $\mathbf{y}_{t+\ell}$, coupling the probability weight assigned to a path segment to the very draw that determines where the path goes next: evaluating even the two-step conditional mean requires integrating the transition probability against the Gaussian density of the next-period state, and the covariance between the two is generically non-zero—an entanglement that propagates at every additional horizon and moment order (see Appendix B.1).

We discuss four tractable approaches that circumvent this obstruction, each with a different computational profile, summarized in Table 1 at the end of the section.

3.1 Simulation-Based Approaches

Standard Monte Carlo The natural baseline is standard Monte Carlo. The initial condition is the pair $(\mathbf{y}_t, \boldsymbol{\pi}_t)$, where $\boldsymbol{\pi}_t = \{\pi_{k,t}\}_{k=1}^{K}$ are the filtered regime probabilities from the Hamilton (1989) filter. Each scenario path is generated by iterating the model forward from $\mathcal{F}_t$, drawing the initial regime from $\boldsymbol{\pi}_t$ and propagating via the reduced-form representation

$$\mathbf{y}_{t+h} = \mathbf{B}_0(s_{t+h}) + \sum_{j=1}^{p} \mathbf{B}_j(s_{t+h})\, \mathbf{y}_{t+h-j} + \mathbf{A}_0(s_{t+h})^{-1}\boldsymbol{\Sigma}(s_{t+h})\, \boldsymbol{\varepsilon}_{t+h}, \tag{20}$$

where $\mathbf{B}_0(k) = \mathbf{A}_0(k)^{-1}\mathbf{c}(k)$ and $\mathbf{B}_j(k) = \mathbf{A}_0(k)^{-1}\mathbf{A}_j(k)$, allowing the regime to evolve stochastically at each step. Relative to existing simulation-based approaches in Markov-switching models that condition on a known regime at the forecast origin (Rubio-Ramírez, Waggoner and Zha, 2005), our procedure integrates over regime uncertainty.

Collect the transition probabilities into the $K \times K$ matrix $\mathbf{P}(\mathbf{y}_t)$, with (j,k) entry $p_{jk}(\mathbf{y}_t)$; under exogenous transitions write $\mathbf{P}$ for the constant version.

Algorithm 1 Standard MC: h-step predictive distribution for MS-VAR

1: **Initialise:** set $\mathbf{y}_{t-j}^{(i)} = \mathbf{y}_{t-j}$, $j = 0, \ldots, p-1$, and draw $s_{t+1}^{(i)} \sim \text{Categorical}(\boldsymbol{\pi}_t)$ for $i = 1, \ldots, N$.
2: **for** $h = 1, \ldots, H$ **do**
3: **Draw innovation:** $\boldsymbol{\varepsilon}_{t+h}^{(i)} \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I})$.
4: **Propagate state:** compute $\mathbf{y}_{t+h}^{(i)}$ via (20) at regime $s_{t+h}^{(i)}$.
5: **Update TP:** compute $\mathbf{P}_{t+h}(\mathbf{y}_{t+h}^{(i)})$ (endogenous) or set $\mathbf{P}_{t+h} = \mathbf{P}$ (exogenous).
6: **Draw next regime:** $s_{t+h+1}^{(i)} \sim \text{Categorical}\Big(\mathbf{e}'_{s_{t+h}^{(i)}}\, \mathbf{P}_{t+h}(\mathbf{y}_{t+h}^{(i)})\Big)$.
7: **end for**
8: **Compute moments:** sample moments of $\{y_{j,t+h}^{(i)}\}_{i=1}^{N}$ at each h.

Under endogenous transitions, the transition matrix in step 5 is evaluated at the simulated $\mathbf{y}_{t+h}^{(i)}$, entangling path weights and path draws as described in Appendix B.1. The root-mean-square error of all sample moment estimates converges at the canonical rate $O(N^{-1/2})$,

independent of the number of variables, lags, or horizons. The per-draw cost is $O(H \cdot d^2)$ where $d = np$.

Two methods reduce the computational burden of Monte Carlo while preserving full generality with respect to the transition function.

Quasi-Monte Carlo (QMC). Standard MC draws pseudo-random uniform points that can cluster unevenly in the integration domain. Replacing them with a *low-discrepancy* Sobol sequence that fills the domain more uniformly yields the same predictive moments at roughly a ten-fold reduction in the number of draws: in practice, $N_q = 10^6$ Sobol paths match the accuracy of $N = 10^7$ standard MC paths at the same computational cost. Implementation requires only swapping the random number generator in Algorithm 1; the loop itself is unchanged. Scrambled Sobol sequences additionally provide independent replicates for standard error estimation at no extra cost. We refer to Appendix B.2 for the Sobol construction, the formal Koksma-Hlawka error bound, and a discussion of how the efficiency gain varies with the integration dimension $D = H(1 + n)$.

Koopman Operator (EDMD). When predictive moments are required at many initial conditions—for instance in historical decompositions or impulse response functions—rerunning Algorithm 1 from scratch at each jumpoff is costly. The Koopman operator approach instead trains a single linear operator $\hat{\mathbf{K}}$ on one-step transitions and then forecasts from any initial condition at negligible marginal cost via iterated matrix-vector products $\hat{\mathbf{K}}^h \boldsymbol{\psi}(\mathbf{y}_0)$, where $\boldsymbol{\psi}$ is a polynomial dictionary of observables. For $n = 5$ variables a degree-3 dictionary with $q = 26$ elements suffices for the first three moments at horizons up to $h = 12$. After a one-time training phase the marginal cost per forecast origin is $O(q^2 H)$, essentially free relative to re-running the simulation. Full convergence analysis, dictionary selection guidance, and numerical validation appear in Appendix B.3.

3.2 An Analytical Shortcut: Probit Transitions

Simulation is unnecessary if one adopts the *probit* link, $p_{jk}(\mathbf{y}) = \Phi(\alpha_{jk} + \beta'_{jk}\mathbf{y})$. The key identity

$$\mathbb{E}[\Phi(a + bX)] = \Phi\left(\frac{a + b\mu}{\sqrt{1 + b^2\sigma^2}}\right), \qquad X \sim \mathcal{N}(\mu, \sigma^2), \tag{21}$$

(proved in Appendix B.4) makes expectations of probit functions against Gaussian measures analytically tractable. The probit specification is observationally near-identical to a logistic (logistic$(y) \approx \Phi(y/1.7)$) but possesses this crucial structural advantage. Computing multi-step conditional moments proceeds in three steps.

Step 1: Path weights. A forward Kalman recursion along each path $\boldsymbol{\kappa}$ delivers the joint Gaussian distribution of intermediate TP-variable realizations. Identity (21) then reduces each path-weight expectation to a rectangular multivariate normal CDF (Appendix B.6).

Step 2: EP correction. The Kalman recursion ignores that conditioning on a known path selects values of the endogenous variables that made each transition likely, biasing component means. An Expectation Propagation (EP) filter corrects for this selection effect at each step using the inverse Mills ratio; under probit transitions the update is exact (Appendix B.7).

Step 3: Moment aggregation. With EP-corrected terminal parameters, the predictive moments follow from the law of total expectation and the marginal predictive density is the exact Gaussian mixture

$$\hat{p}(y) = \sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}} \, \mathcal{N}\left(y;\, \mu_{\boldsymbol{\kappa},v},\, \sigma^2_{\boldsymbol{\kappa},v}\right), \tag{22}$$

evaluated on a grid without kernel smoothing.

Applied to an MS-VAR(6) with $n = 5$, $K = 2$, and probit transitions, this approach delivers all moments in 5.1 s on 32 cores versus 99.2 s for the $N = 10^7$ MC benchmark, with maximum deviations of 0.006 in the mean and 0.002 in the standard deviation across all five variables. The dominant scalability constraint is the K^h path explosion; practical limits are

$h \lesssim 18$ for $K = 2$ and $K \leq 3$ for $h \leq 12$. These limits can be relaxed by pruning paths whose cumulative weights fall below a prescribed threshold, discarding negligible mixture components before they proliferate. A principled implementation of this idea is Kim (1994)'s collapsing step, which merges the K^2 components arising at each period back into K components by matching their first two moments—a moment-matching approximation that keeps the filter tractable at any horizon at the cost of a controlled approximation error. Complexity details and a scalar illustration appear in Appendices B.5 and B.8.

3.3 Comparison of Methods

Table 1 collects the properties of all four approaches. No single method dominates. Standard MC is the universal baseline for any link function, any K, and any h. QMC is a drop-in upgrade—same algorithm, different random number generator—yielding roughly a ten-fold efficiency gain when the integration dimension $D = H(1 + n) \lesssim 100$. The Koopman method has a qualitatively different profile: negligible marginal cost per forecast origin after a one-time training phase, making it the natural choice when moments are required at many initial conditions. The EP filter delivers machine-precision moments and an analytic density but is restricted to probit transitions and moderate horizons.

A side-by-side numerical comparison of all four methods on a simulated MS-VAR with endogenous probit transitions is presented in Appendix C.

3.4 Impulse Responses

In a nonlinear model, impulse responses depend on both the initial condition and the size and sign of the shock, so the linear concept of a fixed impulse response function does not apply. Following Koop, Pesaran and Potter (1996), we define the *generalized impulse response*

Table 1: Methods for multi-step MS-VAR conditional moments: scalability, parallelization, and practical scope. Notation: $d = np$ (companion dimension), $D = H(1 + n)$ (QMC integration dimension), q (dictionary size), $\oplus$ = embarrassingly parallel.

	MC	**QMC (Sobol)**	**Koopman (EDMD)**	**EP Filter**
Cost	$O(N\,h\,d^2)$	$O(N_q\,h\,d^2)$	$O(N_{\text{tr}}\,d) + O(q^2 H)$	$O(K^h(h\,d^2 + (h{-}1)^3))$
Scales with h?	Linear	Linear	Linear (after fit)	Exponential (K^h)
Scales with K?	Linear	Linear	Linear	Exponential (K^h)
Transition function	Any	Any	Any (learned)	Probit: exact; Logit: Smolyak
Moment accuracy	$O(N^{-1/2})$	$O(N_q^{-1}(\log N_q)^D)$	$O(N_{\text{tr}}^{-1/2})$ + bias	Quadrature + interpolation
Density	KDE (kernel bias)	KDE (kernel bias)	Not directly	Exact Gaussian mixture
Deterministic?	No	Yes	No (training)	Yes
Parallel dimension	$\oplus$ draws N	$\oplus$ draws N_q	$\oplus$ training draws	$\oplus$ over K^h paths
Forecast re-use	Full re-run per $\mathbf{y}_0$	Full re-run per $\mathbf{y}_0$	Free: $\hat{\mathbf{K}}^h \boldsymbol{\psi}(\mathbf{y}_0)$	Full re-run per $\mathbf{y}_0$
Practical h limit ($K = 2$)	Unlimited	Unlimited	Unlimited	$h \lesssim 18$
Practical K limit	Any	Any	Any	$K \leq 3$ for $h \leq 12$
Best used when	Universal baseline; any link, any K, any h.	Need 10× fewer draws than MC; $D \lesssim 100$.	Many forecast origins; training cost is amortized.	Moderate h, $K \leq 3$; exact moments or analytic density needed.

function (GIRF) for any scalar functional g of the predictive distribution as

$$\mathrm{GIRF}_g(h, \delta, \mathcal{F}_t) \equiv \mathbb{E}[g(y_{t+h}) \mid \boldsymbol{\varepsilon}_{t+1} = \delta, \mathcal{F}_t] - \mathbb{E}[g(y_{t+h}) \mid \mathcal{F}_t]\,, \tag{23}$$

where δ is the shock vector at impact. Setting $g(y) = y_v$ recovers the familiar mean response for variable v; setting $g(y) = \mathbf{1}(y \leq q)$ yields the response of the predictive CDF at quantile q, and hence the response of $\mathrm{GaR}_{t,h}(\tau)$ to the shock.

Two features of the MS-VAR make GIRFs particularly informative. First, a shock that shifts the mean may simultaneously alter the regime probabilities $\boldsymbol{\pi}_{t+1|t}$, changing the mixture weights and thus the shape of the predictive distribution even holding fixed all structural parameters. A negative shock, for instance, raises $\pi_{2,t+1|t}$ under endogenous transitions, thickening the left tail and worsening GaR by more than a linear model would imply. Second, the response is state-dependent: the same shock generates a larger GaR deterioration when the economy is already near the recession regime than when it is firmly in expansion.

Quantile IRF. Our primary object of interest is the *quantile impulse response function* (QIRF), obtained by setting $g(y) = \mathbf{1}(y \leq q)$ in (23) and solving for the quantile level. Letting $F_j(\cdot;\, h)$ denote the h-step-ahead predictive CDF under scenario $j \in \{\text{base, shock}\}$, the QIRF at quantile τ is

$$\mathrm{QIRF}(\tau, h) = Q_{\text{shock}}(\tau;\, h) - Q_{\text{base}}(\tau;\, h), \tag{24}$$

where $Q_j(\tau;\, h) = F_j^{-1}(\tau;\, h)$. This measures how the τ-quantile trajectory of the forecast distribution shifts in response to the shock, capturing the full distributional effect on tail risks across horizons. We report results for a range of quantiles $\tau \in (0, 1)$, focusing on the difference between the shocked and baseline quantile paths to characterize both central and tail responses.

Computation. Computing (23)–(24) requires evaluating the h-step-ahead predictive distribution at many initial conditions and horizons. We employ two complementary methods; both are described in detail in Appendix D, and we verify below that they deliver virtually identical results.

Forward-recursion (EP) method. The Expectation Propagation filter computes the predictive distribution analytically by enumerating all 2^h possible regime paths from t to $t+h$. Along each path the conditional distribution is Gaussian, and the regime transition probabilities—which depend on the level of the endogenous variables through a probit link—are integrated out via a moment-matching approximation. The h-step predictive density is a mixture of 2^{h+1} Gaussians (blending over the two possible initial regimes), and quantiles are obtained by bisection on the mixture CDF. Applying the precomputed Koopman operator matrix $\hat{\mathbf{K}}^h$ offers an equivalent and computationally efficient route for evaluating (23) at arbitrary $(\mathcal{F}_t, \delta)$ pairs once the operator has been trained.

Monte Carlo method. The MC approach draws N simulations of the model forward from each initial condition, using *paired* random draws: the baseline and shocked simulations share the same sequence of standardized innovations and regime-transition uniforms, so all sampling variability is common to both scenarios. At horizon h, the QIRF is computed by *quantile matching*—comparing the empirical τ-quantile of the shocked cross-sectional distribution to that of the baseline distribution—exactly paralleling definition (24). Standard MC and QMC can also be applied by conditioning the first innovation draw on δ and simulating freely thereafter.[2]

[2]An alternative MC approach takes the τ-quantile of the pairwise differences $Y_{h,i}^{\text{shock}} - Y_{h,i}^{\text{base}}$. Because $Q_\tau(X-Y) \neq Q_\tau(X) - Q_\tau(Y)$ for nonlinearly related random variables, this does not correspond to definition (24) and yields artificially narrow quantile bands in the presence of regime switching. We discuss this further in Appendix D.

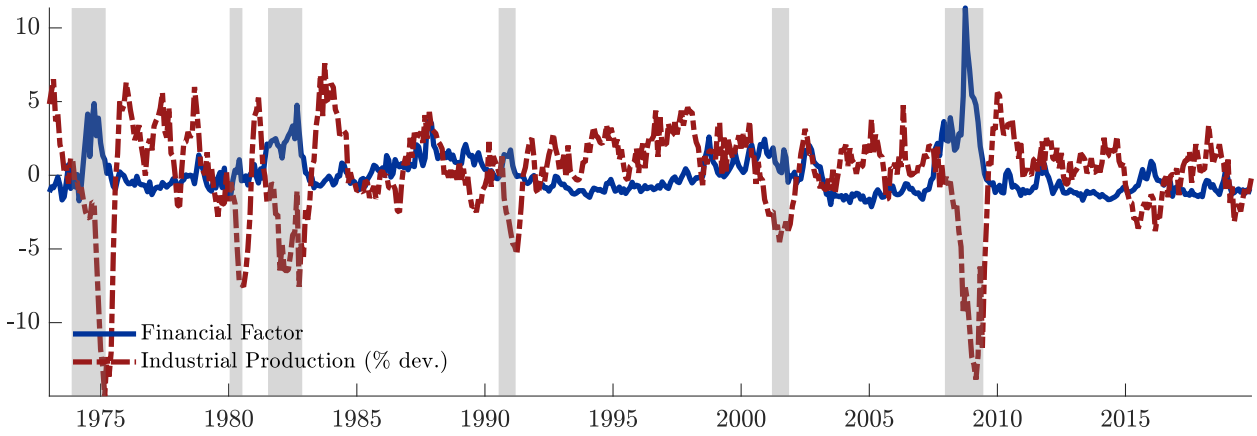
4 Quantifying Downside Risk

We now present an empirical implementation of the Markov-Switching VAR framework to quantify downside risk and macroeconomic fragility in the United States. We begin with a parsimonious bivariate MS-VAR using industrial production and a financial factor to demonstrate the model's ability to capture asymmetric predictive distributions and construct operational measures of upside and downside risk. During the 2007-2008 Global Financial Crisis, the model quantifies the build-up of macroeconomic and financial fragility, with even modest bad-regime probabilities generating substantial asymmetries in the predictive densities.

4.1 A Look at the Data

The dataset includes monthly observations of industrial production and a financial factor. Industrial production represents key real-sector output and economic activity data. For the financial factor, a distinct Dynamic Factor Model (DFM) is employed to analyze monthly financial variables. The information set for financial factors consists of a parsimonious selection of representative variables that encompass credit and financial conditions, non-financial vulnerabilities, and macro-financial imbalances. Specifically, the dataset includes observations for the VXO, a volatility index derived from the S&P 100, which reflects market uncertainty; the excess bond premium (EBP), as outlined by Gilchrist and Zakrajšek (2012) and Favara et al. (2016), which measures credit market risk; the TED spread, defined as the difference between the 3-month LIBOR rate and the 3-month Treasury bill rate, indicating interbank lending stress; and the CBILL spread, which is the difference between the 3-month financial commercial paper rate and the 3-month Treasury bill rate, reflecting short-term funding conditions in the financial sector.

Figure 7: Macroeconomic Activity and Financial Conditions



Note: The top panel plots the cyclical component of IP growth. The bottom panel plots corporate borrowing spreads. All figures are in percentage deviations. Shaded areas denote NBER recessions.

4.2 Calibration of Endogenous Transition Probabilities

The endogenous transition probabilities $p_{jk}(\mathbf{y}_t)$ introduce additional parameters — the intercepts α_{jk} and slope coefficients β_{jk} — that govern how current conditions affect the probability of switching regimes. We discuss three approaches to calibrating these.

Joint MLE. The most direct approach is to include the transition parameters in the full model likelihood and estimate them jointly with the VAR coefficients via the EM algorithm or direct numerical optimization of the Hamilton (1989) filter likelihood. This is internally consistent but can be computationally demanding in high-dimensional systems, and identification may be weak if regime switches are rare in the estimation sample.

Matching a vulnerability indicator. An alternative is to construct an external vulnerability indicator v_t — a scalar index that summarizes current financial or macroeconomic stress — and calibrate α_{jk} and β_{jk} so that the implied transition probabilities $p_{jk}(\mathbf{y}_t)$ track v_t closely. In practice this amounts to a probit or logit regression of observed regime transitions on v_t,

where the regime sequence is either taken from an external chronology (e.g. NBER dates) or inferred from a first-pass exogenous MS-VAR. The advantage is that the vulnerability indicator can aggregate information from variables outside the VAR, such as credit spreads or survey measures of uncertainty.

Matching business cycle moments. A third approach calibrates the transition parameters to match salient features of the business cycle, such as the average duration of expansions and recessions, the unconditional frequency of recession, and the degree of asymmetry in exit probabilities. Given a target vector of moments $\mathbf{m}^*$, one solves

$$\min_{\alpha_{jk},\,\beta_{jk}} \left\| \mathbf{m}(\alpha_{jk}, \beta_{jk}) - \mathbf{m}^* \right\|_W^2, \tag{25}$$

where $\mathbf{m}(\cdot)$ are the model-implied moments and W is a weighting matrix. Because the ergodic regime probabilities and expected durations are smooth functions of the transition parameters under the logistic or probit link, the minimization in (25) is well-behaved and can be solved quickly. This approach is especially useful when one wants to impose that the model replicates known stylized facts — for instance, that recessions last on average four quarters and that the probability of entering recession rises sharply when financial conditions deteriorate.

In our empirical application we follow the second and third approaches in combination: we first construct a vulnerability indicator from financial and macroeconomic variables, then calibrate α_{jk} and β_{jk} to match both the path of the indicator and the empirical duration of NBER recession and expansion episodes.

4.3 A Bivariate Example

We start by studying a simple bivariate MS-VAR to show that this framework is flexible enough to model downside risk. We are mainly interested in constructing the predictive

density (7). Our specification is parsimonious with six lags, and monthly data on two observables: industrial production and financial factor. These are the same variables entering the endogenous transition probabilities (18).

We use Sims and Zha (1998) priors with hyperparameters $\lambda_1 = 1$ and $\lambda_3 = 1$. Here λ_1 controls the overall shrinkage of the lag coefficients and λ_3 controls the lag decay, allowing the influence of lagged values to diminish appropriately over time.

Table 2 shows estimated parameters of the reduced-form representation (20) for the bivariate model.

Table 2: Estimated Parameters: Reduced Form

		Industrial Production		Financial Factor	
Parameter	Description	Regime 1	Regime 2	Regime 1	Regime 2
$d(s)$	Constant	0.07	-0.03	-0.33	0.39
$\sigma(s)$	Std dev	0.88	1.82	0.41	1.33
$\sigma_{12}(s)$	Covariance	0.00	0.00	0.00	0.10
$\sum_{j=1}^{p} b_j(s)$	Persistence	0.84	0.84	0.93	0.92
$\bar{Y}(s)$	Cond. Mean	0.65	-5.75	-0.27	4.34

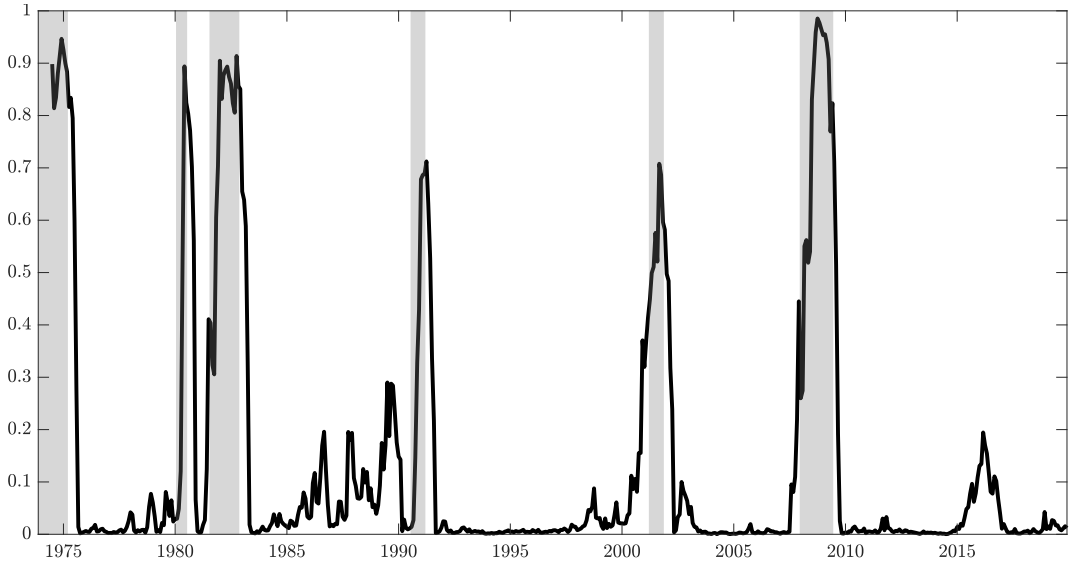
In contrast with other work using MS-VAR we are not interested in obtaining filtered or smoothed estimates of the model-implied regimes. Instead we focus on constructing simulation-based metrics that characterize the components of the predictive distribution.

Figure 8 shows the evolution of the one-year-ahead probability of the bad regime conditional on time t information: $\Pr(s_{t+12} = 2 \mid \mathcal{F}_t, \hat{\boldsymbol{\theta}})$. Our model-implied simulations generate reasonable estimates of the underlying evolution of the unobserved regimes thanks to our calibration procedure.

Our framework allows us to jointly construct predictive distributions for all variables in the MS-VAR model. This is one key advantage relative to approaches such as quantile regressions that can only recover the predictive distribution for a single variable of interest.

Figure 9 plots the quantiles of the one-year-ahead forecast distribution of industrial

Figure 8: Average Simulated Probability of Bad Regime



Note:

production in the top panel and the financial factor in the bottom panel. The black line represents the 50th percentile, the dashed red line is the 10th percentile, and the dotted blue line is the 90th percentile.

4.4 Macroeconomic Asymmetries

To gauge the asymmetry in the predictive densities, we calculate the difference between selected quantiles. Upside risk is measured by the difference between the 90th and 50th percentiles; downside risk by the difference between the 50th and 10th percentiles.

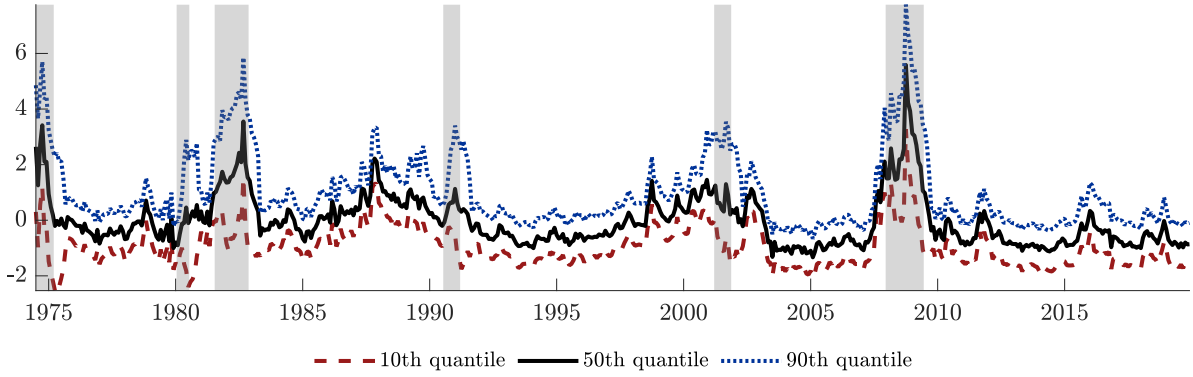
Figure 10 shows the asymmetry measures for the financial factor (top panel) and industrial production (bottom panel).

4.5 Build-up of Risk: 2007–2008

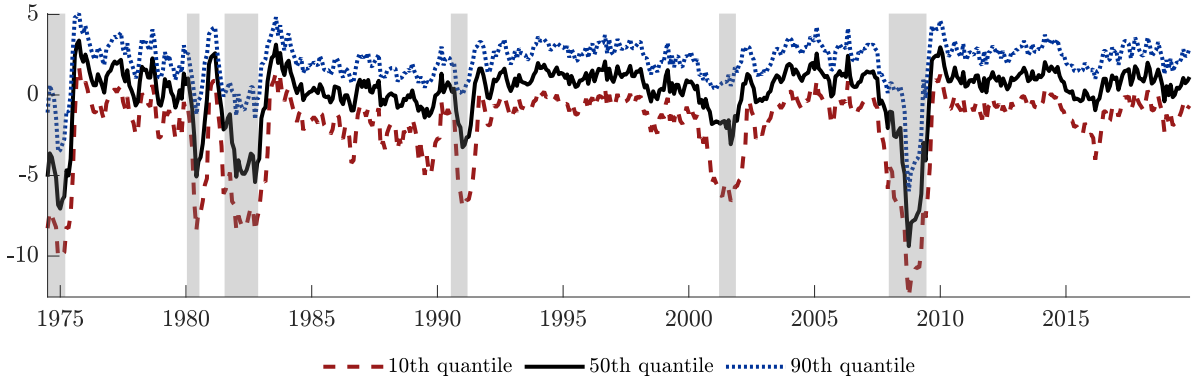
On August 16, 2007, the Federal Open Market Committee (FOMC) held an emergency conference call and decided to lower the discount rate by 50 basis points, from 6.25% to 5.75%,

Figure 9: Conditional Quantiles

(a) Avg. Financial Factor Over Next 12 Months



(b) Avg. Industrial Production Over Next 12 Months



Note:

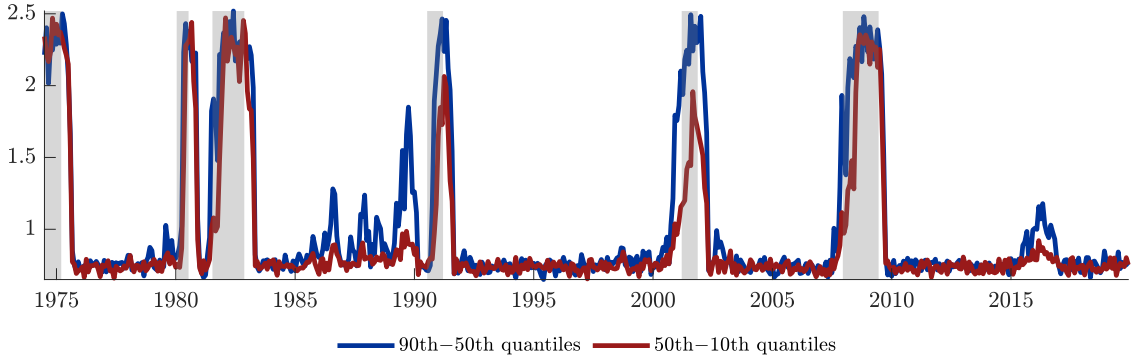
in response to financial market stress caused by the unfolding subprime mortgage crisis.

> *"Financial market conditions have deteriorated, and tighter credit conditions and increased uncertainty have the potential to restrain economic growth going forward. In these circumstances, although recent data suggest that the economy has continued to expand at a moderate pace, the Federal Open Market Committee judges that the downside risks to growth have increased appreciably."*
>
> *August 16, 2007 — Conference Call*

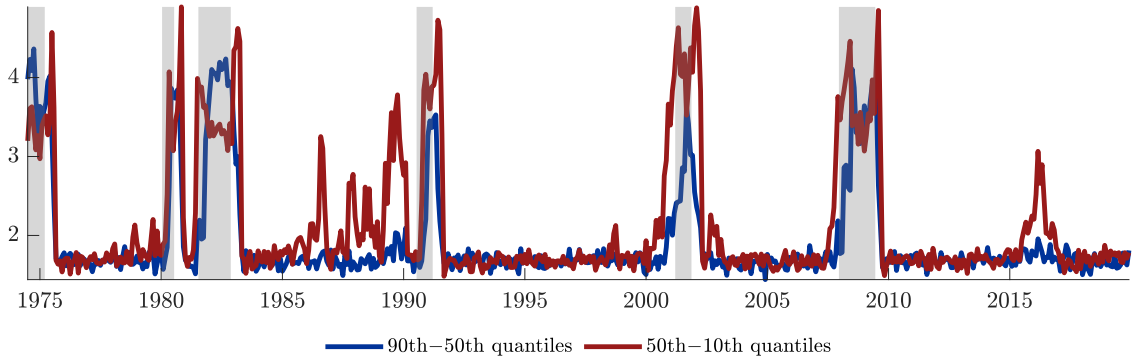
Figure 11 illustrates how our MS-VAR model quantifies the extent of downside risk at the

Figure 10: Macroeconomic Asymmetries

(a) Asymmetries of Avg. Financial Factor Over Next 12 Months



(b) Asymmetries of Avg. Industrial Production Over Next 12 Months



Note:

time of the FOMC emergency conference call. The left panel shows conditional distributions for industrial production growth in the good (green-dotted) and bad (red-dashed) regimes, while the right panel shows conditional distributions for the financial factor. Solid black lines are the 12-months-ahead predictive density obtained by averaging the regime-specific distributions over the forecast horizon, weighted at each step by the probability of being in each regime:

$$f\left(\frac{1}{H}\sum_{h=1}^{H}\mathbf{y}_{t+h}\,\middle|\,\mathcal{F}_t\right) = \frac{1}{H}\sum_{h=1}^{H}\pi_{1,t+h|t}\,f\left(\frac{1}{H}\sum_{h=1}^{H}\mathbf{y}_{t+h}\,\middle|\,\{s_{t+h}=1\}_{h=1}^{H},\mathcal{F}_t\right) + \frac{1}{H}\sum_{h=1}^{H}\pi_{2,t+h|t}\,f\left(\frac{1}{H}\sum_{h=1}^{H}\mathbf{y}_{t+h}\,\middle|\,\{s_{t+h}=2\}_{h=1}^{H},\mathcal{F}_t\right), \tag{26}$$

which is the horizon-averaged counterpart of the one-step predictive mixture (7).

Conditional one-year-ahead distributions are consistent with industrial production and the financial factor being in the good regime, but with a bad-regime probability of 11%. Such a small but non-negligible probability is sufficient to create downside risks, generating a noticeable left skew for industrial production and right skew for the financial factor.

It is important to note that the regime-specific distributions are not necessarily Gaussian, as they depend on the probability of transitioning across regimes. To illustrate, consider the one-step-ahead distribution conditional on $s_{t+1} = 1$. The components $f(\mathbf{y}_{t+1} \mid s_{t+1} = 1, s_t = k)$ are Gaussian given our conditionally linear VAR structure, but the mixture weight $p_{12}(\mathbf{y}_t)$ is nonlinear and state-dependent:

$$f(\mathbf{y}_{t+1} \mid s_{t+1} = 1, \mathcal{F}_t) = \Big(1 - p_{12}(\mathbf{y}_t)\Big)\, f(\mathbf{y}_{t+1} \mid s_{t+1} = 1, s_t = 1) + p_{21}(\mathbf{y}_t)\, f(\mathbf{y}_{t+1} \mid s_{t+1} = 1, s_t = 2)\,. \tag{27}$$

Figure 12 quantifies the magnitude of downside risk during the GFC, which we refer to as *fragility*. We compute $\int_{-\infty}^{b} f(y_{i,t+h} \mid \mathcal{F}_t)\, dy$ for industrial production and $\int_{b}^{\infty} f(y_{i,t+h} \mid \mathcal{F}_t)\, dy$ for the financial factor, across horizons $h = 1, \ldots, 12$. The threshold b corresponds to the realized value of each variable during NBER recessions: $b = -3.8$ for IP and $b = 1.3$ for FF,

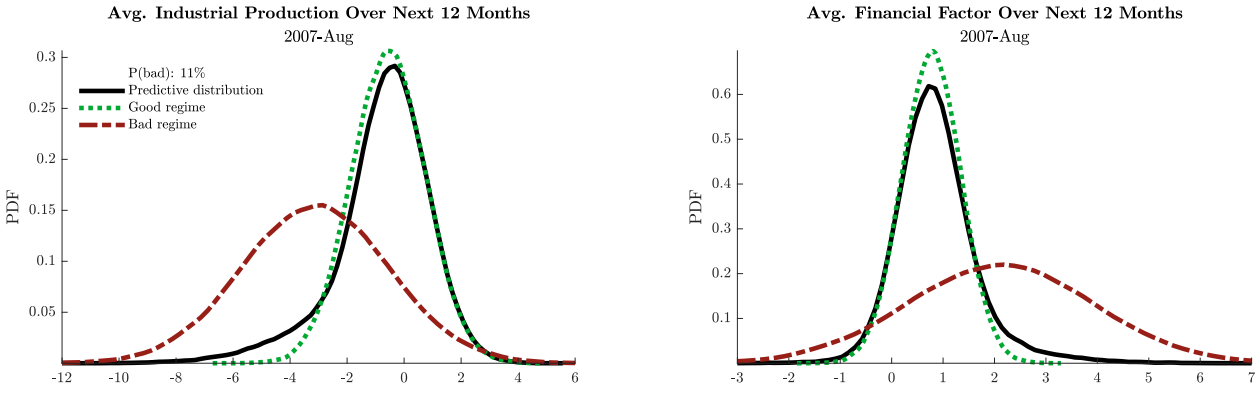
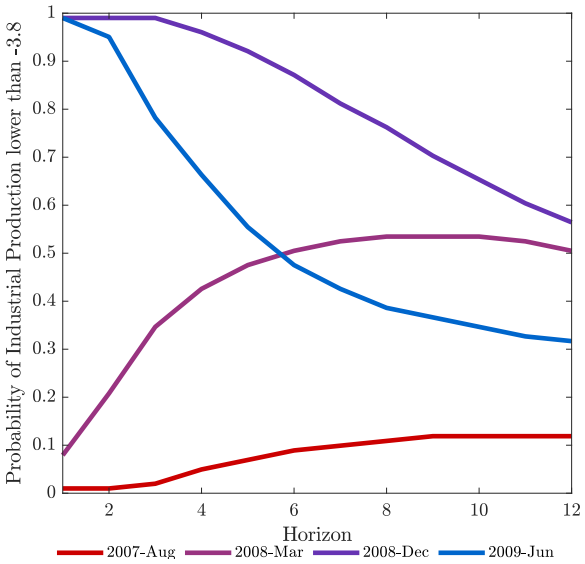


Figure 11: Macroeconomic and Financial Risks During the Global Financial Crisis

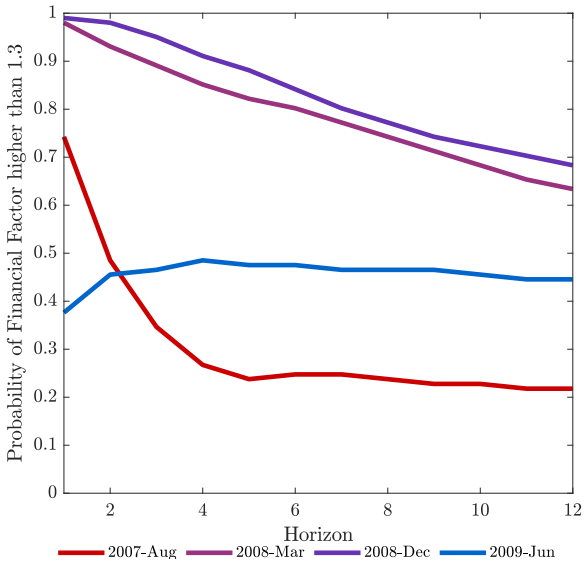
Note:

Figure 12: Fragility during the GFC

(a) Industrial production fragility



(b) Financial factor fragility



Note:

set at the mode of the data in NBER-identified recession periods. Fragility thus measures the probability of outcomes worse than their historical average during recessions.

The left panel of Figure 12 shows IP fragility for four different periods. In August 2007, the model detected relatively modest fragility at short horizons, rising steadily over the one-year outlook. The right panel shows that fragility was markedly more severe for financial variables: tighter financing conditions were expected to reach recession levels with a probability above 70% in the near term, settling to about 25% within a year. This is consistent with the FOMC's assessment at the time of their emergency meeting.

In subsequent months, macroeconomic and financial fragility rose. By March 2008, when the bad-regime probability had jumped to nearly 100%, IP fragility was on the rise while financial fragility moderated somewhat, likely reflecting the Federal Reserve's actions to shore up financial conditions by rescuing Bear Stearns. By December 2008, both measures had jumped to nearly 100%, with the model interpreting the data as implying financial conditions

for the coming year would remain tighter than during historical recessions.

5 The Tightening of Lending Conditions

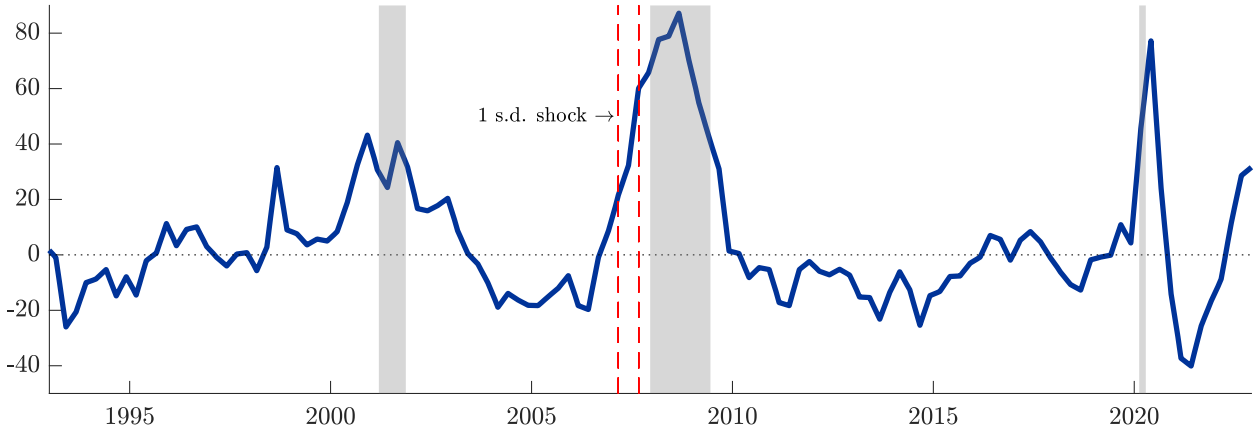
In this section we extend our model to highlight its flexibility as a tool for structural analysis. We study the transmission of a tightening of lending conditions at the onset of the GFC, focusing on developments around August 2007.

5.1 Data and Identification

We extend our MS-VAR model to include six monthly U.S. macroeconomic variables in $\mathbf{y}_t$, ordered as follows: core PCE inflation, unemployment rate, industrial production, SLOOS, financial factor, and the shadow federal funds rate. As in Section 4.3, we maintain the assumption that the transition probabilities are a function of (IP_t, FF_t) and keep the calibration of transition probability coefficients as described in Section 4.2.

We estimate the structural form (19) from March 1992 to December 2019; the sample is limited by the availability of the SLOOS. The structural identification follows a recursive

Figure 13: U.S. Aggregate Lending Standards



Note:

ordering where the SLOOS is placed after the real variables (inflation, unemployment rate, industrial production) but before the financial factor and the shadow federal funds rate, as in Broadbent et al. (2024). This is the most conservative approach to extracting the endogeneity from the SLOOS, purging it from factors that may simultaneously affect credit demand, while maintaining a reasonable monetary policy reaction function. By construction, monetary policy does not respond contemporaneously to the SLOOS, since that survey is released with a delay.

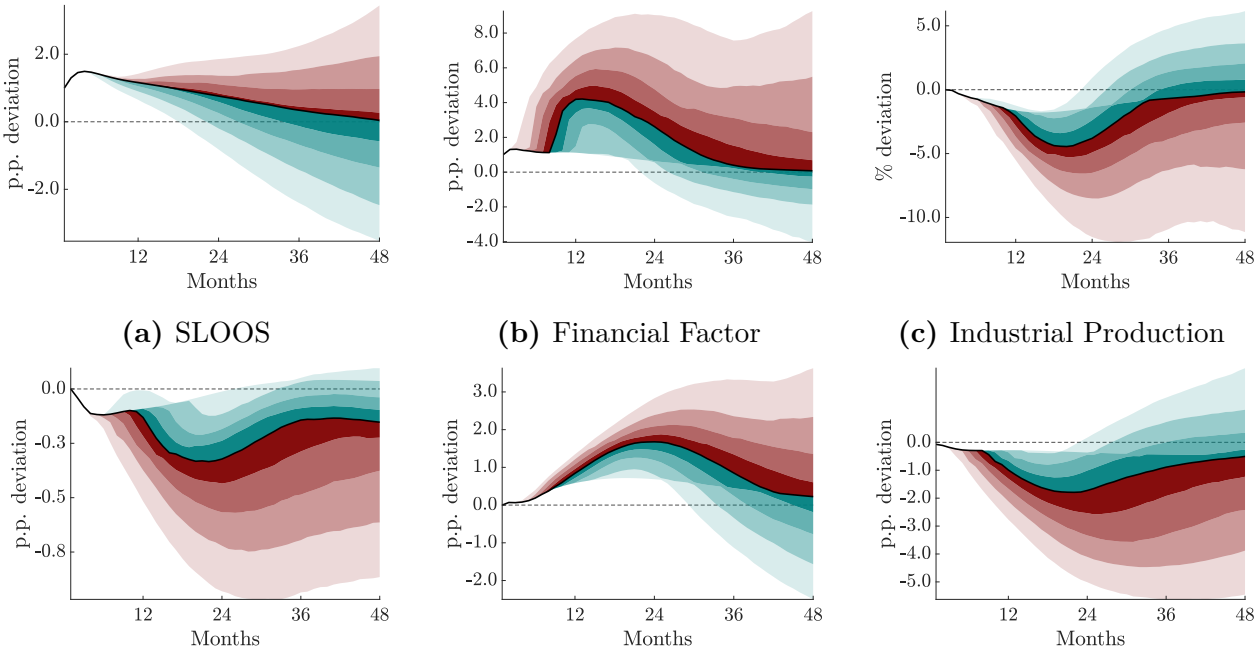
5.2 Impulse Responses

Figure 14 shows impulse responses to a one-standard-deviation shock to the SLOOS, representing a tightening of lending conditions in August 2007. Responses are presented at the posterior mean of the parameters; distribution bands (shaded areas) reflect uncertainty about the regime path. As described by Algorithm 1, keeping parameters fixed, the full simulation of the model over the forecasting horizon delivers a distribution of potential paths depending on initial conditions, estimated parameters, and transition probabilities.

The responses are constructed following the GIRF definition (23), where the effect at each horizon h is the difference between the shocked and baseline paths. Under endogenous transitions, the regime path s_{t+h} depends not only on the sequence of regime draws $\{z_{t+h}\}$ but also on the shock $\boldsymbol{\varepsilon}_{t+1} = \delta$, since the shock affects the financial factor and industrial production, which in turn enter the transition probabilities $p_{12}(\mathbf{y}_t)$ and $p_{21}(\mathbf{y}_t)$.

A one-standard-deviation increase in the SLOOS, if observed at the onset of the GFC in August 2007, would generate a large deterioration of the economic outlook. The tightening of lending conditions would induce an increase in financial stress and depress activity. Effects peak between one and two years after the shock, with risks tilted to the downside. The shock is deflationary and raises the unemployment rate, putting monetary policy in an uncomfortable

position. The model implies an easing of the stance, though the range of potential paths shows that either higher or lower federal funds rates are possible in the medium to long run, depending on the evolution of economic and financial conditions.

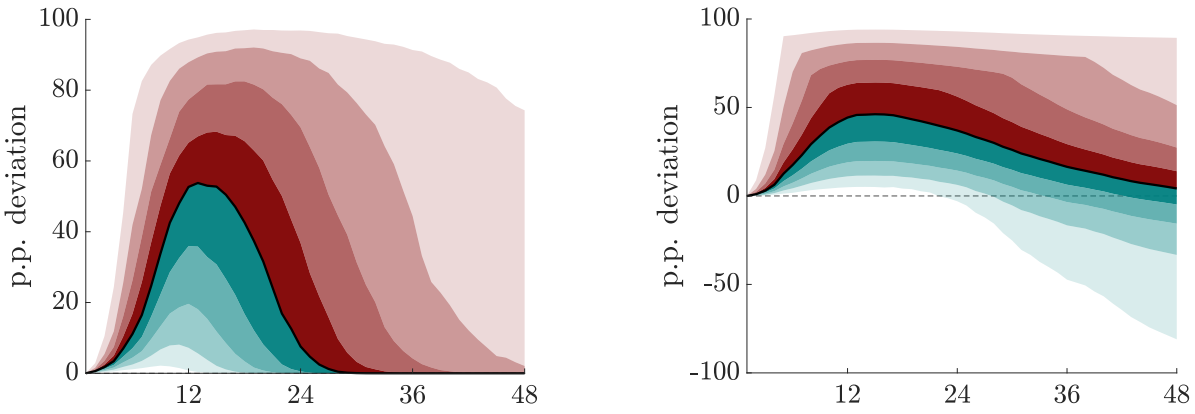


(a) SLOOS **(b)** Financial Factor **(c)** Industrial Production

(d) Core PCE Inflation **(e)** Unemployment Rate **(f)** Federal Funds Rate

Figure 14: Responses to tighter lending conditions in August 2007

Figure 15 shows impulse responses for the transition probability $p_{12}(\mathbf{y}_t)$ (left panel) and the bad-regime probability $\pi_{2,t+h|t}$ (right panel). The deterioration of industrial production and the financial factor raises the probability of switching from the good to the bad regime beyond what the baseline path implies. At its height, this probability increases by about 50 percentage points, with downside scenarios where switching is near certain. Consequently, the bad-regime probability is also substantially larger in the medium run, though eventual policy responses can lead to paths where the bad-regime probability falls below the unconditional baseline in the long run.



(a) $p_{12}(\mathbf{y}_t)$: Good→Bad **(b)** $\pi_{2,t+h|t}$: Bad-regime probability

Figure 15: Responses of transition probabilities to tighter lending conditions in August 2007

5.3 Downside Risk During the GFC

Figure 16 further illustrates the downside risks of a tightening of lending conditions and the difficult position for monetary policy at the onset of the GFC. The figures show joint predictive distributions of the 12-months-ahead average of pairs of variables; blue dashed contours represent the baseline unconditional forecasts and red solid contours represent the shocked distributions following a one-standard-deviation tightening of lending conditions.

Panels 16a and 16b give a clear picture of the downside risk in real and financial conditions. A tightening of bank lending conditions shifts the joint distribution of SLOOS and industrial production in the south-east direction (16a), indicating higher mean SLOOS and lower mean industrial production, with the distribution skewed toward even lower industrial production realizations. Similarly, the shock shifts the joint distribution of SLOOS and the financial factor to the north-east (16b), indicating higher financial stress with a pronounced right skew.

Panel 16c illustrates the dual mandate of the Federal Reserve. The tightening shifts the joint distribution of core inflation and the unemployment rate in the north-west direction, indicating a deflationary stance coupled with higher unemployment, with risks skewed toward

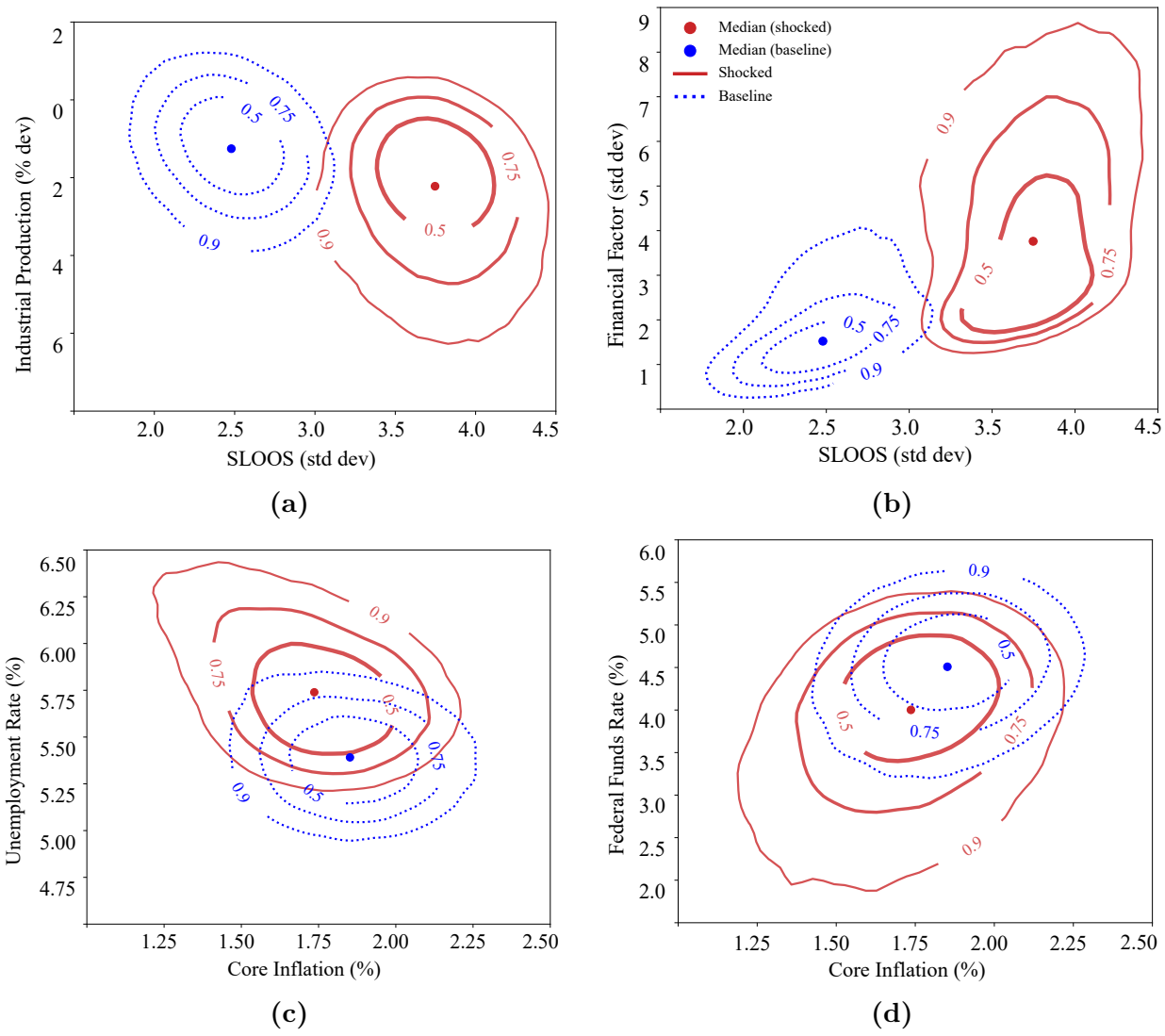


Figure 16: One-year-ahead joint densities following a SLOOS shock in August 2007

scenarios that exacerbate this trade-off. As shown in Panel 16d, the shock shifts the joint distribution of core inflation and the federal funds rate to the south-west, with distributions skewed toward a substantial easing of the monetary policy stance.

6 Conclusion

References

Adrian, Tobias, Federico Grinberg, Nellie Liang, Sheheryar Malik, and Jie Yu. 2022. "The Term Structure of Growth-at-Risk." *American Economic Journal: Macroeconomics*, 14(3): 283–323.

Adrian, Tobias, Nina Boyarchenko, and Domenico Giannone. 2019. "Vulnerable Growth." *American Economic Review*, 109(4): 1263–1289.

Adrian, Tobias, Nina Boyarchenko, and Domenico Giannone. 2021. "Multimodality In Macrofinancial Dynamics." *International Economic Review*, 62(2): 861–886.

Adrian, Tobias, Nina Boyarchenko, Francesco Furno, Domenico Giannone, Leonardo Iania, Michele Lenza, Francesca Loria, Cecilia Melo Fernandes, and Sergio Sola. 2025. "Macro Risk." Unpublished manuscript.

Bianchi, Francesco. 2016. "Methods for measuring expectations and uncertainty in Markov-switching models." *Journal of Econometrics*, 190(1): 79–99.

Bocola, Luigi. 2016. "The Pass-Through of Sovereign Risk." *Journal of Political Economy*, 124(4): 879–926.

Broadbent, Elijah, Huberto M. Ennis, Tyler J. Pike, and Horacio Sapriza. 2024. "Bank Lending Standards and the U.S. Economy."

Caldara, Dario, Chiara Scotti, and Molin Zhong. 2021. "Macroeconomic and Financial Risks: A Tale of Mean and Volatility." Board of Governors of the Federal Reserve System (U.S.) International Finance Discussion Papers 1326.

Camacho, Maximo, Gabriel Perez-Quiros, and Pilar Poncela. 2018. "Markov-switching dynamic factor models in real time." *International Journal of Forecasting*, 34(4): 598–611.

Chavleishvili, Sulkhan, and Simone Manganelli. 2023. "Forecasting and Stress Testing with Quantile Vector Autoregression." *Journal of Applied Econometrics*, 1–.

Delle Monache, Davide, Andrea De Polis, and Ivan Petrella. 2024. "Modeling and Forecasting Macroeconomic Downside Risk." *Journal of Business & Economic Statistics*, 42(3): 1010–1025.

De Polis, Andrea, Leonardo Melosi, and Ivan Petrella. 2024. "The Taming of the Skew: The Taming of the Skew: Asymmetric Inflation Risk and Monetary Policy." *CEPR Discussion Paper*, DP19760.

Diebold, Francis X., and Glenn D. Weinbach. 1994. "Regime Switching with Time-Varying Transition Probabilities." *Business Cycles: Theory and Empirical Methods*, 283–313.

Doz, Catherine, Laurent Ferrara, and Pierre-Alain Pionnier. 2020. "Business Cycle Dynamics after the Great Recession." Paris School of Economics (PSE) PSE Working Paper.

Favara, Giovanni, Simon Gilchrist, Kurt F Lewis, and Egon Zakrajšek. 2016. "Recession Risk and the Excess Bond Premium." Board of Governors of the Federal Reserve System (US).

Fernández-Villaverde, Jesús, Samuel Hurtado, and Galo Nuno. 2023. "Financial frictions and the wealth distribution." *Econometrica*, 91(3): 869–901.

Filardo, Andrew J. 1994. "Business Cycle Phases and Their Transitional Dynamics." *Journal of Business and Economic Statistics*, 12(3): 299–308.

Gallant, A Ronald, Peter E Rossi, and George Tauchen. 1993. "Nonlinear Dynamic Structures." *Econometrica*, 61(4): 871–907.

Giglio, Stefano, Bryan Kelly, and Seth Pruitt. 2016. "Systemic risk and the macroeconomy: An empirical evaluation." *Journal of Financial Economics*, 119(3): 457–471.

Gilchrist, Simon, and Egon Zakrajšek. 2012. "Credit Spreads and Business Cycle Fluctuations." *American Economic Review*, 102(4): 1692–1720.

Gourio, François. 2012. "Disaster Risk and Business Cycles." *American Economic Review*, 102(6): 2734–66.

Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica*, 57(2): 357–384.

Harding, Don, and Adrian Pagan. 2002. "Dissecting the Cycle: A Methodological Investigation." *Journal of Monetary Economics*, 49(2): 365–381.

Joe, Stephen, and Frances Y Kuo. 2008. "Constructing Sobol sequences with better two-dimensional projections." *SIAM Journal on Scientific Computing*, 30(5): 2635–2654.

Karamé, Frédéric. 2010. "Impulse-Response Functions in Markov-Switching Structural Vector Autoregressions: A Step Further." *Economics Letters*, 106(3): 162–165.

Karamé, Frédéric. 2015. "Asymmetries and Markov-Switching Structural VAR." *Journal of Economic Dynamics and Control*, 53: 85–102.

Karlsson, Sune, Stepan Mazur, and Hoang Nguyen. 2023. "Vector Autoregression Models with Skewness and Heavy Tails." *Journal of Economic Dynamics and Control*, 146: 104580. Working Paper originally issued as Örebro University Working Paper No. 8/2021.

Kim, Chang-Jin. 1994. "Dynamic linear models with Markov-switching." *Journal of econometrics*, 60(1-2): 1–22.

Kole, Erik, and Dick van Dijk. 2023. "Moments, Shocks and Spillovers in Markov-Switching VAR Models." *Journal of Econometrics*, 236(2): 105474.

Koop, Gary, M. Hashem Pesaran, and Simon M. Potter. 1996. "Impulse Response Analysis in Nonlinear Multivariate Models." *Journal of Econometrics*, 74(1): 119–147.

Montes-Galdón, Carlos, and Eva Ortega. 2022. "Skewed SVARs: Tracking the Structural Sources of Macroeconomic Tail Risks." *Essays in Honour of Fabio Canova (Advances in Econometrics, Vol. 44A)*, 177–210. Emerald Publishing Limited.

Plagborg-Møller, Mikkel, Lucrezia Reichlin, Giovanni Ricco, and Thomas Hasenzagl. 2020. "When is growth at risk?" *Brookings Papers on Economic Activity*, 2020(1): 167–229.

Rubio-Ramírez, Juan F., Daniel F. Waggoner, and Tao Zha. 2005. "Markov-Switching Structural Vector Autoregressions: Theory and Application." *Federal Reserve Bank of Atlanta Working Paper*.

Salgado, Sergio, Fatih Guvenen, and Nicholas Bloom. 2019. "Skewed business cycles." National Bureau of Economic Research.

Sims, Christopher A., and Tao Zha. 1998. "Bayesian Methods for Dynamic Multivariate Models." *International Economic Review*, 39(4): 949–968.

Sims, Christopher A., and Tao Zha. 2006. "Were There Regime Switches in U.S. Monetary Policy?" *American Economic Review*, 96(1): 54–81.

Timmermann, Allan. 2000. "Moments of Markov Switching Models." *Journal of Econometrics*, 96(1): 75–111.

A Proofs and Derivations for Section 2

A.1 Derivation of Path Mean and Variance

We derive expressions (5)–(6) by induction on h. Fix a regime path $\boldsymbol{\kappa} = (k_1, \ldots, k_h) \in \{1, 2\}^h$ and condition throughout on $\boldsymbol{\kappa}$ and $\mathcal{F}_t$.

Base case ($h = 1$). Conditional on $s_{t+1} = k_1$, the model gives $y_{t+1} = \mu_{k_1} + \phi_{k_1} y_t + \sigma_{k_1} \eta_{t+1}$, so

$$m_{(k_1)} = \mu_{k_1} + \phi_{k_1} y_t, \qquad v_{(k_1)} = \sigma_{k_1}^2,$$

which matches (5)–(6) at $h = 1$ (empty products equal one).

Inductive step. Suppose (5)–(6) hold for paths of length $h - 1$. For a path of length h, write $y_{t+h} = \mu_{k_h} + \phi_{k_h} y_{t+h-1} + \sigma_{k_h} \eta_{t+h}$, where η_{t+h} is independent of $\mathcal{F}_{t+h-1}$ and of the path up to step $h - 1$. By the inductive hypothesis, $y_{t+h-1} \mid (k_1, \ldots, k_{h-1}), \mathcal{F}_t$ is Gaussian with mean $m_{(k_1,\ldots,k_{h-1})}$ and variance $v_{(k_1,\ldots,k_{h-1})}$. Taking conditional expectation and variance:

$$\begin{aligned}
m_{\boldsymbol{\kappa}} &= \mu_{k_h} + \phi_{k_h}\, m_{(k_1,\ldots,k_{h-1})} \\
&= \mu_{k_h} + \phi_{k_h} \left[\Big(\prod_{\ell=1}^{h-1} \phi_{k_\ell} \Big) y_t + \sum_{j=1}^{h-1} \Big(\prod_{\ell=j+1}^{h-1} \phi_{k_\ell} \Big) \mu_{k_j} \right] \\
&= \Big(\prod_{\ell=1}^{h} \phi_{k_\ell} \Big) y_t + \sum_{j=1}^{h-1} \Big(\prod_{\ell=j+1}^{h} \phi_{k_\ell} \Big) \mu_{k_j} + \mu_{k_h},
\end{aligned}$$

which is (5) since the last term corresponds to $j = h$ with an empty product equal to one. For the variance, using independence of η_{t+h} from y_{t+h-1}:

$$\begin{aligned} v_{\boldsymbol{\kappa}} &= \phi_{k_h}^2 \, v_{(k_1,\ldots,k_{h-1})} + \sigma_{k_h}^2 \\ &= \phi_{k_h}^2 \sum_{j=1}^{h-1} \Big(\prod_{\ell=j+1}^{h-1} \phi_{k_\ell}^2 \Big) \sigma_{k_j}^2 + \sigma_{k_h}^2 \\ &= \sum_{j=1}^{h-1} \Big(\prod_{\ell=j+1}^{h} \phi_{k_\ell}^2 \Big) \sigma_{k_j}^2 + \sigma_{k_h}^2, \end{aligned}$$

which is (6) since the last term corresponds to $j = h$ with an empty product equal to one. This completes the induction. □

A.2 Proof of Proposition 1

The three expressions follow from the law of total expectation and variance applied to the mixture (4). Throughout, write $\pi_k \equiv \pi_{k,t+1|t}$ and, conditional on $s_{t+1} = k$, $y_{t+1} \sim \mathcal{N}(\mu_k + \phi_k y_t,\, \sigma_k^2)$.

Conditional mean.

$$\mathbb{E}[y_{t+1} \mid \mathcal{F}_t] = \sum_{k=1}^{2} \pi_k (\mu_k + \phi_k y_t) = \bar{\mu} + \bar{\phi}\, y_t.$$

Conditional variance. By the law of total variance:

$$\mathbb{V}[y_{t+1} \mid \mathcal{F}_t] = \sum_{k=1}^{2} \pi_k \sigma_k^2 + \sum_{k=1}^{2} \pi_k \Big[(\mu_k + \phi_k y_t) - (\bar{\mu} + \bar{\phi}\, y_t) \Big]^2 = \sum_{k=1}^{2} \pi_k \sigma_k^2 + \sum_{k=1}^{2} \pi_k d_k^2.$$

Conditional skewness. Writing $y_{t+1} - \bar{\mu} - \bar{\phi}\, y_t = \sigma_k \eta_{t+1} + d_k$ and expanding the cube, using $\mathbb{E}[\eta_{t+1}] = 0$, $\mathbb{E}[\eta_{t+1}^2] = 1$, $\mathbb{E}[\eta_{t+1}^3] = 0$:

$$\mathbb{E}\Big[(y_{t+1} - \bar{\mu} - \bar{\phi}\, y_t)^3 \mid s_{t+1} = k, \mathcal{F}_t\Big] = 3\sigma_k^2 d_k + d_k^3.$$

Averaging over k with weights π_k and dividing by $\mathbb{V}[y_{t+1} \mid \mathcal{F}_t]^{3/2}$ gives (13). □

A.3 Derivation of Multi-Step Conditional Moments

We derive expressions (14)–(16) from the mixture representation (7). Throughout, write $w_{\boldsymbol{\kappa}} = \Pr(s_{t+1} = k_1, \ldots, s_{t+h} = k_h \mid \mathcal{F}_t)$ for the path weights, which satisfy $\sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}} = 1$, and recall that conditional on a path $\boldsymbol{\kappa}$, $y_{t+h} \mid \boldsymbol{\kappa}, \mathcal{F}_t \sim \mathcal{N}(m_{\boldsymbol{\kappa}}, v_{\boldsymbol{\kappa}})$ with mean and variance given by the recursions (5)–(6).

Conditional mean. By the law of total expectation, marginalizing over all 2^h paths:

$$\mathbb{E}[y_{t+h} \mid \mathcal{F}_t] = \sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}}\, \mathbb{E}[y_{t+h} \mid \boldsymbol{\kappa}, \mathcal{F}_t] = \sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}}\, m_{\boldsymbol{\kappa}} \equiv \bar{m}_h,$$

which is (14).

Conditional variance. By the law of total variance:

$$\begin{aligned}
\mathbb{V}[y_{t+h} \mid \mathcal{F}_t] &= \mathbb{E}[\mathbb{V}[y_{t+h} \mid \boldsymbol{\kappa}, \mathcal{F}_t]] + \mathbb{V}[\mathbb{E}[y_{t+h} \mid \boldsymbol{\kappa}, \mathcal{F}_t]] \\
&= \sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}}\, v_{\boldsymbol{\kappa}} + \sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}} (m_{\boldsymbol{\kappa}} - \bar{m}_h)^2 \\
&= \sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}}\, v_{\boldsymbol{\kappa}} + \sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}}\, d_{\boldsymbol{\kappa}}^2,
\end{aligned}$$

where $d_{\boldsymbol{\kappa}} \equiv m_{\boldsymbol{\kappa}} - \bar{m}_h$. This is (15).

Third central moment. Write $y_{t+h} - \bar{m}_h = (y_{t+h} - m_{\boldsymbol{\kappa}}) + d_{\boldsymbol{\kappa}}$ and expand the cube:

$$(y_{t+h} - \bar{m}_h)^3 = (y_{t+h} - m_{\boldsymbol{\kappa}})^3 + 3(y_{t+h} - m_{\boldsymbol{\kappa}})^2 d_{\boldsymbol{\kappa}} + 3(y_{t+h} - m_{\boldsymbol{\kappa}}) d_{\boldsymbol{\kappa}}^2 + d_{\boldsymbol{\kappa}}^3.$$

Taking the conditional expectation given path $\boldsymbol{\kappa}$ and $\mathcal{F}_t$, and using the fact that $y_{t+h} \mid \boldsymbol{\kappa}, \mathcal{F}_t$ is Gaussian with mean $m_{\boldsymbol{\kappa}}$ and variance $v_{\boldsymbol{\kappa}}$, so that its within-path third central moment is zero and its within-path second central moment is $v_{\boldsymbol{\kappa}}$:

$$\mathbb{E}\Big[(y_{t+h} - \bar{m}_h)^3 \mid \boldsymbol{\kappa}, \mathcal{F}_t\Big] = 0 + 3\, v_{\boldsymbol{\kappa}}\, d_{\boldsymbol{\kappa}} + 0 + d_{\boldsymbol{\kappa}}^3 = 3\, v_{\boldsymbol{\kappa}}\, d_{\boldsymbol{\kappa}} + d_{\boldsymbol{\kappa}}^3.$$

Averaging over paths with weights $w_{\boldsymbol{\kappa}}$:

$$\mathbb{E}\Big[(y_{t+h} - \bar{m}_h)^3 \mid \mathcal{F}_t\Big] = \sum_{\boldsymbol{\kappa}} w_{\boldsymbol{\kappa}} \Big(3\, v_{\boldsymbol{\kappa}}\, d_{\boldsymbol{\kappa}} + d_{\boldsymbol{\kappa}}^3\Big),$$

which is (16). Note that the within-path skewness term is identically zero because each path is conditionally Gaussian: skewness at the mixture level arises entirely from heterogeneity *across* paths, through the cross term $3\, v_{\boldsymbol{\kappa}}\, d_{\boldsymbol{\kappa}}$ and the between term $d_{\boldsymbol{\kappa}}^3$, exactly mirroring the one-step decomposition in Proposition 1. □

B Proofs and Derivations for Section 3

B.1 Cross-Term Integrals under Endogenous Transitions

We derive the cross-term integrals that arise at each moment order and establish that the covariance implied by (**??**) is generically nonzero.

Conditional mean at $h = 2$. Starting from the law of iterated expectations, for each pair (k_1, k_2) the required integral is

$$I^{(1)}_{k_1k_2} = \int p_{k_1k_2}(y) \cdot y \cdot \phi(y; m_{k_1}, \sigma^2_{k_1})\, dy = \mathrm{Cov}(p_{k_1k_2}(y_{t+1}),\, y_{t+1} \mid k_1) + m_{k_1} \cdot \bar{p}_{k_1k_2}, \tag{B-1}$$

where $m_{k_1} = \mu_{k_1} + \phi_{k_1} y_t$ and $\bar{p}_{k_1k_2} = \mathbb{E}[p_{k_1k_2}(y_{t+1}) \mid k_1]$. Under exogenous TP the covariance vanishes, giving the closed form $I^{(1)} = p_{k_1k_2} \cdot m_{k_1}$. Under endogenous TP with monotone increasing $p_{k_1k_2}$ ($\beta_{k_1k_2} > 0$), the function assigns higher probability to $y_{t+1} > m_{k_1}$, generating a strictly positive covariance; a decreasing function gives strictly negative covariance. The covariance is zero if and only if $\beta_{k_1k_2} = 0$ for all (k_1, k_2).

Higher-order cross-terms. For the variance and skewness the relevant integrals are

$$I^{(r)}_{k_1k_2} = \int p_{k_1k_2}(y) \cdot y^r \cdot \phi(y; m_{k_1}, \sigma^2_{k_1})\, dy = \mathrm{Cov}(p_{k_1k_2}(y),\, y^r \mid k_1) + \mathbb{E}[y^r \mid k_1] \cdot \bar{p}_{k_1k_2}, \quad r = 2, 3. \tag{B-2}$$

Under the logistic link each $I^{(r)}$ is a logistic-normal integral with no closed form. Under the probit link, identity (21) evaluates $\bar{p}_{k_1k_2}$ in closed form, but the covariance terms still require the backward GH recursion. At horizon h the cross-terms compound via nested expectations over products of transition probabilities.

B.2 Quasi-Monte Carlo: Sobol Construction and Convergence

Construction. The Sobol sequence is a family of base-2 digital nets. In each dimension j, the i-th point is constructed using direction numbers $v_{j,1}, \ldots, v_{j,m}$ (integers satisfying a linear recurrence over $\mathbb{F}_2$) via binary XOR:

$$u_j^{(i)} = \frac{i_1\, v_{j,1} \oplus i_2\, v_{j,2} \oplus \cdots \oplus i_m\, v_{j,m}}{2^m}, \tag{B-3}$$

where $i = \sum_k i_k 2^{k-1}$ is the binary expansion of i. The direction numbers are chosen so that all s-dimensional projections satisfy the (t, s)-net property: every elementary interval of volume 2^{t-m} contains exactly 2^t points. Using the Gray code, each successive point requires only a single XOR, making generation $O(1)$ per point.

Koksma-Hlawka bound and dimension effects. The integration error satisfies

$$\left| \frac{1}{N} \sum_{i=1}^{N} g(F^{-1}(\mathbf{u}^{(i)})) - \int_{[0,1]^D} g(F^{-1}(\mathbf{u}))\, d\mathbf{u} \right| \leq V[g \circ F^{-1}] \cdot D_N^*, \tag{B-4}$$

where $V[\cdot]$ is the Hardy-Krause variation and $D_N^* = O(N^{-1}(\log N)^D)$ is the Sobol discrepancy. The asymptotic bound appears unfavourable for large D, but the Koksma-Hlawka inequality is a worst-case upper bound. In practice, for MS-VAR integrands the *effective dimension* is substantially lower than D: the first forecast step dominates the variation and the remaining steps contribute progressively less. For $D \lesssim 100$ the practical Sobol gain is approximately one order of magnitude at realistic N, consistent with the numerical evidence in Joe and Kuo (2008).

Randomised QMC. Scrambled Sobol sequences (random digital shift plus Owen-type nested scrambling) preserve the low-discrepancy property while making R independent

replicates available for variance estimation:

$$\hat{\sigma}^2_{\text{QMC}} = \frac{1}{R(R-1)} \sum_{r=1}^{R} (\hat{\mu}_r - \bar{\mu})^2, \qquad \bar{\mu} = \frac{1}{R} \sum_{r=1}^{R} \hat{\mu}_r. \tag{B-5}$$

With $R = 16$ scrambled replicates of N_q/R points each, a confidence interval for the moment estimate is obtained at essentially the same cost as a single N_q run.

Interaction with endogenous TP. Under exogenous transitions the mapping $F^{-1} : [0,1]^D \to \mathbb{R}^D$ is fixed across paths, keeping the Hardy-Krause variation small. Under endogenous transitions, the transition CDF at step h depends on the simulated $\mathbf{y}^{(i)}_{t+h}$ and varies across paths, increasing the effective variation. This mildly reduces the Sobol advantage at short horizons, but the gain remains substantial because the innovation draws (n of the $1+n$ dimensions per step) enter through the fixed Φ^{-1} transformation.

B.3 Koopman Operator: Convergence and Dictionary Selection

Convergence of EDMD. Let $\mathcal{K} : L^2(\rho) \to L^2(\rho)$ be the Koopman operator, where ρ is the stationary distribution of $(\mathbf{y}_t, s_t)$, and let $\mathbf{K}^*_q = \Pi_q \mathcal{K} \Pi_q$ be its best rank-q approximation. Under β-mixing of $(\mathbf{y}_t, s_t)$ at geometric rate, the EDMD estimator

$$\hat{\mathbf{K}} = (\mathbf{\Psi}_t' \mathbf{\Psi}_t)^{-1} \mathbf{\Psi}_t' \, \mathbf{\Psi}_{t+1} \,, \tag{B-6}$$

where $\mathbf{\Psi}_t \in \mathbb{R}^{N_{\text{tr}} \times q}$ collects the dictionary evaluated at N_{tr} one-step training pairs, satisfies

$$\|\hat{\mathbf{K}}_q - \mathbf{K}^*_q\|_F = O_p\!\left(\sqrt{q/N_{\text{tr}}}\right). \tag{B-7}$$

The total h-step forecasting error decomposes as

$$\|\hat{\mathbf{K}}_q^h \boldsymbol{\psi}(\mathbf{y}_0) - \mathcal{K}^h \boldsymbol{\psi}(\mathbf{y}_0)\| \leq \underbrace{h\|\hat{\mathbf{K}}_q\|^{h-1}\|\hat{\mathbf{K}}_q - \mathbf{K}_q^*\|_F}_{\text{estimation error } O_p(h\sqrt{q/N_{\text{tr}}})} + \underbrace{\|\mathbf{K}_q^{*h} - \mathcal{K}^h\|_{\text{op}}\|\boldsymbol{\psi}\|}_{\text{truncation bias } \delta_q^{(h)}}. \tag{B-8}$$

The estimation error grows with h as errors accumulate across matrix applications; stability is enforced by projecting eigenvalues of $\hat{\mathbf{K}}_q$ onto the unit disk.

Dictionary selection. By the Stone-Weierstrass theorem, polynomial dictionaries are dense in $L^2(\rho)$ for any compactly supported ρ, so the truncation bias $\delta_q^{(h)}$ vanishes as $q \to \infty$. For smooth MS dynamics, a degree-3 dictionary ($q \approx 26$ for $n = 5$) suffices for the first three moments at $h \leq 12$. The dictionary should be built from the current observation $\mathbf{y}_t$ only, not the full companion state $\mathbf{Y}_t \in \mathbb{R}^{np}$: lags enter linearly so Koopman eigenfunctions are smooth in $\mathbf{y}_t$ alone, and including all $d = np$ dimensions inflates q by a factor of p^2 without proportional accuracy gains.

Multi-horizon forecasting. Once fitted, forecasts at all horizons $h = 1, \ldots, H$ cost $O(q^2 \log H)$ via successive squaring of $\hat{\mathbf{K}}_q$, then $O(qH)$ per initial condition via matrix-vector products. This amortises the $O(N_{\text{tr}} q^2)$ training cost over all forecast origins.

B.4 The Probit-Gaussian Identity

Proposition 2 (Probit-Gaussian convolution)**.** *Let* $X \sim \mathcal{N}(\mu, \sigma^2)$*. Then* $\mathbb{E}[\Phi(a + bX)] = \Phi\Big((a + b\mu)/\sqrt{1 + b^2\sigma^2}\Big)$.

Proof. Introduce independent $Z \sim \mathcal{N}(0, 1)$. Then $\mathbb{E}[\Phi(a + bX)] = P(Z \leq a + bX) = P(Z - bX \leq a)$. Since $Z - bX \sim \mathcal{N}(-b\mu, 1 + b^2\sigma^2)$, standardising gives the result. □

B.5 Closed-Form Moments for MS-AR under Probit Transitions

We derive the backward GH recursion and closed-form raw moment formulas for the univariate MS-AR described in Section 3.2.

Selection bias. A naive recursion sets $\mathbb{E}[y_{t+h} \mid \boldsymbol{\kappa}, y_t] = \mu_{k_h} + \phi_{k_h} m_{\boldsymbol{\kappa}}^{(h-1)}$. This fails whenever $\beta_{jk} \neq 0$: conditioning on regime k_h at step h selects values of y_{t+h-1} through $p_{k_{h-1},k_h}(y_{t+h-1})$, a Heckman-type selection effect.

Proposition 3 (Selection bias under state-dependent transitions)**.** *Under probit transitions with $\beta_{k_{h-1},k_h} \neq 0$, the naive path mean $\tilde{m}^{(h)} = \mu_{k_h} + \phi_{k_h} \tilde{m}^{(h-1)}$ satisfies*

$$m_{\boldsymbol{\kappa}}^{(h)} - \tilde{m}_{\boldsymbol{\kappa}}^{(h)} = \phi_{k_h} \frac{\mathrm{Cov}(p_{k_{h-1}k_h}(y_{t+h-1}),\, y_{t+h-1} \mid k_{h-1}, \mathcal{F}_t)}{\mathbb{E}[p_{k_{h-1}k_h}(y_{t+h-1}) \mid k_{h-1}, \mathcal{F}_t]} \neq 0. \tag{B-9}$$

Proof. Conditioning on path $\boldsymbol{\kappa}$ selects y_{t+h-1} with weight $p_{k_{h-1}k_h}(y_{t+h-1})$, so $\mathbb{E}[y_{t+h-1} \mid \boldsymbol{\kappa}] = \mathbb{E}[p_{k_{h-1}k_h}(y_{t+h-1}) \cdot y_{t+h-1} \mid k_{h-1}]/\mathbb{E}[p_{k_{h-1}k_h}(y_{t+h-1}) \mid k_{h-1}]$. Decomposing the numerator as in (B-1) and subtracting $\tilde{m}^{(h)}$ yields the stated expression; nonzero β_{k_{h-1},k_h} implies nonzero covariance by the monotone function argument in Appendix B.1. □

Backward GH recursion. The correct approach maintains, for each path $\boldsymbol{\kappa}$ and $r \in \{0, 1, 2, 3\}$, the function

$$G_\ell^{(r)}(y) = \mathbb{E}\left[y_{t+h-1}^r \prod_{j=\ell}^{h-1} \Phi\big(\alpha_{k_j,k_{j+1}} + \beta_{k_j,k_{j+1}}\, y_{t+j}\big) \;\middle|\; y_{t+\ell-1} = y,\ \boldsymbol{\kappa} \right], \tag{B-10}$$

initialised as $G_h^{(r)}(y) = y^r$. The backward recursion for $\ell = h-1, \ldots, 1$ is

$$G_\ell^{(r)}(y) = \frac{1}{\sqrt{\pi}} \sum_{q=1}^{Q} w_q\, \Phi(a_\ell + b_\ell\, \xi_q(y))\ G_{\ell+1}^{(r)}(\xi_q(y))\,, \tag{B-11}$$

where $\xi_q(y) = \mu_{k_\ell} + \phi_{k_\ell} y + \sigma_{k_\ell}\sqrt{2}\, z_q$ and (z_q, w_q) are Q-point GH nodes and weights.

Proposition 4 (Backward GH recursion)**.** *Under probit transitions, $G_\ell^{(r)}$ satisfies* (B-11).

Proof. By the Markov property and the regime-k_ℓ Gaussian transition density: $G_\ell^{(r)}(y) = \int \Phi(a_\ell + b_\ell y')\, G_{\ell+1}^{(r)}(y')\, \phi(y'; \mu_{k_\ell} + \phi_{k_\ell} y, \sigma_{k_\ell}^2)\, dy'$. The change of variables $u = (y' - \mu_{k_\ell} - \phi_{k_\ell} y)/(\sigma_{k_\ell}\sqrt{2})$ converts this to $\frac{1}{\sqrt{\pi}} \int \Phi(a_\ell + b_\ell \xi(u; y))\, G_{\ell+1}^{(r)}(\xi(u; y))\, e^{-u^2}\, du$, which is in standard GH form; applying the Q-point rule gives (B-11). □

Raw moment accumulation. Let $\tilde{g}_{\boldsymbol{\kappa}}^{(r)} = \Phi(\alpha_{s_0,k_1} + \beta_{s_0,k_1} y_0) \cdot G_1^{(r)}(y_0)$. With $\mu \equiv \mu_{k_h}$, $\phi \equiv \phi_{k_h}$, $\sigma \equiv \sigma_{k_h}$:

Proposition 5 (Closed-form raw moments under probit transitions)**.**

$$\mathbb{E}[y_{t+h} \cdot \mathbf{1}_{\boldsymbol{\kappa}} \mid y_t, s_t] = \mu\, \tilde{g}_{\boldsymbol{\kappa}}^{(0)} + \phi\, \tilde{g}_{\boldsymbol{\kappa}}^{(1)}, \tag{B-12}$$

$$\mathbb{E}\left[y_{t+h}^2 \cdot \mathbf{1}_{\boldsymbol{\kappa}} \mid y_t, s_t\right] = (\mu^2 + \sigma^2)\, \tilde{g}_{\boldsymbol{\kappa}}^{(0)} + 2\mu\phi\, \tilde{g}_{\boldsymbol{\kappa}}^{(1)} + \phi^2\, \tilde{g}_{\boldsymbol{\kappa}}^{(2)}, \tag{B-13}$$

$$\mathbb{E}\left[y_{t+h}^3 \cdot \mathbf{1}_{\boldsymbol{\kappa}} \mid y_t, s_t\right] = (\mu^3 + 3\mu\sigma^2)\, \tilde{g}_{\boldsymbol{\kappa}}^{(0)} + (3\mu^2\phi + 3\phi\sigma^2)\, \tilde{g}_{\boldsymbol{\kappa}}^{(1)} + 3\mu\phi^2\, \tilde{g}_{\boldsymbol{\kappa}}^{(2)} + \phi^3\, \tilde{g}_{\boldsymbol{\kappa}}^{(3)}. \tag{B-14}$$

Proof. Write $y_{t+h} = \mu + \phi y_{t+h-1} + \sigma\varepsilon$ with $\varepsilon \perp \mathbf{1}_{\boldsymbol{\kappa}}$, $\mathbb{E}[\varepsilon^{2j+1}] = 0$, $\mathbb{E}[\varepsilon^2] = 1$. The first moment follows immediately. For the second, expand y_{t+h}^2 and use $\mathbb{E}[\varepsilon^2] = 1$ and $\mathbb{E}[y_{t+h-1}\varepsilon \cdot \mathbf{1}] = 0$. For the third, expand y_{t+h}^3 via the binomial theorem and apply the odd-moment conditions; the $3\phi\sigma^2$ coefficient of $\tilde{g}^{(1)}$ arises from $3\mathbb{E}[\phi y_{t+h-1} \cdot \sigma^2\varepsilon^2 \cdot \mathbf{1}]$. □

Summing (B-12)–(B-14) over all 2^h paths and applying standard cumulant relations yields the conditional mean, variance, and skewness.

B.6 Multivariate Normal CDF Reduction for Probit Path Weights

Proposition 6 (Probit path weights as a multivariate normal probability)**.** *Under the joint Gaussian $(Z_1, \ldots, Z_{h-1}) \sim \mathcal{N}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$ from the forward Kalman recursion, the path-weight*

expectation satisfies

$$\mathbb{E}\left[\prod_{t=1}^{h-1}\Phi(u_t(Z_t))\right] = P(W_t \le a_t,\ \forall\, t)\,, \quad \mathbf{W} \sim \mathcal{N}(-B'\boldsymbol{\mu}_Z,\ B'\boldsymbol{\Sigma}_Z B + \mathbf{I}_{h-1}), \tag{B-15}$$

with $B = [\boldsymbol{b}_{k_1,k_2}, \ldots, \boldsymbol{b}_{k_{h-1},k_h}]$.

Proof. Each factor $\Phi(u_t(Z_t)) = P(\eta_t \le u_t(Z_t))$ for independent $\eta_t \sim \mathcal{N}(0,1)$. Let $W_t = \eta_t - \boldsymbol{b}'_{k_t,k_{t+1}} Z_t$; then $\prod_t \Phi(u_t(Z_t)) = \mathbf{1}(W_t \le a_t, \forall t)$, and $\boldsymbol{W} \sim \mathcal{N}(-B'\boldsymbol{\mu}_Z,\ B'\boldsymbol{\Sigma}_Z B + \mathbf{I}_{h-1})$ follows from independence of $\boldsymbol{\eta}$ and $\boldsymbol{Z}$. □

B.7 EP State Update under Probit Transitions

Proposition 7 (EP update for probit selection)**.** *At each step $t < h$ along path $\boldsymbol{\kappa}$, with predicted state $\mathbf{Y}_t^{\text{pred}} \sim \mathcal{N}(\boldsymbol{\mu}_t^{\text{pred}}, \boldsymbol{\Sigma}_t^{\text{pred}})$ and linear index $L_t \sim \mathcal{N}(\ell_t, v_t)$, the EP posterior given $\{L_t > 0\}$ has mean and covariance*

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_t^{\text{pred}} + \mathbf{K}_g \frac{r}{\sqrt{v_t}}, \tag{B-16}$$

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_t^{\text{pred}} - \frac{r(\tilde{a}+r)}{v_t}\,\mathbf{K}_g\,\mathbf{K}_g', \tag{B-17}$$

where $\ell_t = \alpha_{k_t,k_{t+1}} + \boldsymbol{b}_t'\mathbf{C}\boldsymbol{\mu}_t^{\text{pred}}$, $v_t = 1 + \boldsymbol{b}_t'(\mathbf{C}\boldsymbol{\Sigma}_t^{\text{pred}}\mathbf{C}')\boldsymbol{b}_t$, $\tilde{a} = \ell_t/\sqrt{v_t}$, $r = \phi(\tilde{a})/\Phi(\tilde{a})$, and $\mathbf{K}_g = \boldsymbol{\Sigma}_t^{\text{pred}}\mathbf{C}'\boldsymbol{b}_t$.

Proof. The truncated-normal moments of $L_t \mid L_t > 0$ are $\mathbb{E}[L_t \mid L_t > 0] = \ell_t + \sqrt{v_t}\,r$ and $\text{Var}[L_t \mid L_t > 0] = v_t(1 - r(\tilde{a}+r))$. Writing $\mathbf{Y}_t = \boldsymbol{\mu}_t^{\text{pred}} + v_t^{-1}\mathbf{K}_g(L_t - \ell_t) + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \perp L_t$, conditioning on $L_t > 0$ gives $\mathbb{E}[\mathbf{Y}_t \mid L_t > 0] = \boldsymbol{\mu}_t^{\text{pred}} + \mathbf{K}_g r/\sqrt{v_t}$, yielding (B-16). The covariance reduction $v_t - \text{Var}[L_t \mid L_t > 0] = v_t r(\tilde{a}+r)$ propagated through $\mathbf{K}_g$ gives (B-17). □

B.8 Complexity of the EP Filter

The dominant scalability constraint is the K^h growth in paths. For $K = 2$:

h	K^h	Approx. storage	EP time (32 cores, est.)
6	64	< 1 KB	$\ll 1$ s
12	4,096	~ 1 MB	~ 6 s
18	262,144	~ 60 MB	~ 5 min
24	16,777,216	~ 4 GB	infeasible

All path-level computations (Kalman recursion, EP correction, weight computation) are embarrassingly parallel over the K^h paths with no inter-worker communication. For $K = 3$, $h = 12$: $3^{12} = 531{,}441$ paths, still feasible. For $K = 4$, $h = 12$: $4^{12} \approx 16.8 \times 10^6$; memory becomes the binding constraint.

C MS-VAR Example with Probit Transition

C.1 Model Design

We consider an $n = 5$, $p = 6$ Markov-switching VAR with $K = 2$ regimes and a probit transition function. The companion state has dimension $d = np = 30$. The two regimes represent a *bad state* (regime 1) and a *good state* (regime 2), with opposing intercepts and high within-regime persistence.

Regime intercepts.

$$\boldsymbol{c}_1 = +3.0\,(0.5,\ -0.3,\ 0.8,\ -0.1,\ 0.4)',$$

$$\boldsymbol{c}_2 = -3.0\,(0.5,\ -0.3,\ 0.8,\ -0.1,\ 0.4)',$$

so the two regimes push the system in exactly opposite directions. This strong separation is the primary driver of the bimodal forecast distributions observed at $h = 12$.

Lag dynamics. A base lag-1 matrix B_1 captures cross-sectional dynamics. Regime 1 sets $\Phi_{1,\ell} = 1.1\, B_1 \times 0.4^{\ell-1}$, producing moderate persistence with geometrically decaying higher-order lags. Regime 2 uses $\Phi_{2,1} = 1.15\, B_1 + 0.08\, I_n$ and $\Phi_{2,\ell} = \Phi_{2,1} \times 0.4^{\ell-1}$ for $\ell \geq 2$, so the good state has slightly higher persistence and stronger own-variable feedback.

Innovation covariances. Both regimes share the same diagonal scale vector (1.0, 0.8, 1.2, 0.9, 1.1). Regime 1 multiplies by 0.4 (tight innovations). Regime 2 multiplies by 0.5 (slightly more volatile) and adds small off-diagonal Cholesky entries linking variables 1–2 and 1–3, introducing mild cross-variable shock propagation.

Transition probabilities. The TP variables are $(Y_{t,1}, Y_{t,2})$, the first two elements of the companion state (i.e. current $y_{t,1}$ and $y_{t,2}$). The intercepts $\alpha_{:,1} = (3.0, -1.5)'$ and $\alpha_{:,2} = (-3.0, 1.5)'$ create very high within-regime persistence: when the system is in the bad state ($y_{t,1} \ll 0$, $y_{t,2} \gg 0$), the probit index for staying in regime 1 is large positive, making switching very unlikely. The opposite holds in the good state. This calibration generates strongly bimodal forecast distributions by horizon $h = 12$.

Initial condition. We set $\boldsymbol{Y}_0 = \mathbf{0}$ (companion state at the origin) and $s_0 = 1$ (starting in the bad state), with forecast horizon $h = 12$.

C.2 Numerical Results

Table C-1 reports the estimated mean, standard deviation, and skewness for all five variables at $h = 12$. The benchmark is standard MC with $N = 10^7$ draws (runtime 99.2 s). QMC uses $N_q = 10^6$ Sobol draws (10× fewer, runtime 13.7 s). The Koopman operator is learned from

$N_{\text{tr}} = 5 \times 10^5$ one-step pairs and forecast with $N_k = 10^6$ draws (runtime 13.5 s). The Forward EP filter enumerates all $2^{12} = 4096$ paths; weight computation (11-dimensional rectangular normal CDF) takes 4.7 s and the EP correction pass takes 0.4 s, totalling 5.1 s on 32 cores.

Table C-1: Estimated moments at $h = 12$, $n = 5$, $p = 6$, $K = 2$, probit transitions. MC ($N = 10^7$, 99.2 s) is the benchmark.

	Variable 1			**Variable 2**			**Variable 3**		
Method	Mean	Std	Skew	Mean	Std	Skew	Mean	Std	Skew
MC (10^7)	−6.3036	7.5803	0.185	2.4741	2.2880	−0.619	−5.5837	4.1706	0.808
QMC (10^6)	−6.2950	7.5828	0.182	2.4763	2.2877	−0.620	−5.5809	4.1696	0.805
Koopman	−6.3004	7.5817	0.185	2.4742	2.2846	−0.617	−5.5839	4.1728	0.806
Forward EP	−6.2977	7.5820	0.184	2.4753	2.2872	−0.619	−5.5813	4.1717	0.806
Max err. vs MC:									
QMC	0.0085	0.0025	0.003	0.0022	0.0003	0.001	0.0028	0.0011	0.003
Koop	0.0032	0.0014	0.000	0.0001	0.0033	0.002	0.0001	0.0023	0.001
EP	0.0059	0.0017	0.001	0.0012	0.0008	0.000	0.0024	0.0011	0.001

	Variable 4			**Variable 5**		
Method	Mean	Std	Skew	Mean	Std	Skew
MC (10^7)	0.1689	2.1733	−0.180	−3.8447	4.5659	0.144
QMC (10^6)	0.1696	2.1711	−0.179	−3.8408	4.5669	0.142
Koopman	0.1675	2.1738	−0.183	−3.8439	4.5671	0.144
Forward EP	0.1689	2.1723	−0.181	−3.8412	4.5672	0.143

Accuracy. All four methods agree closely across all three moment types. For the mean, the largest deviation from the MC benchmark is 0.009 (QMC, variable 1); for the standard deviation it is 0.003 (Koopman, variable 2). The Forward EP filter achieves mean error ≤ 0.006 and standard deviation error ≤ 0.002 across all variables, without drawing a single random number. Skewness estimates are consistent in sign and magnitude across all methods: positive for variables 1, 3, 5 (reflecting the net bad-state direction of those intercept components) and negative for variables 2 and 4 (where the bad-state intercept is negative and dominates the

mean).

Computational cost. The Forward EP filter runs in 5.1 s total on 32 cores, a $\approx 19\times$ wall-clock speedup over the MC benchmark (99.2 s), while producing a smooth analytic density rather than a kernel estimate. QMC achieves comparable moment accuracy at 10^6 draws in 13.7 s. The Koopman operator requires 13.5 s for both learning and forecasting.

C.3 Forecast Density Comparison

Figure C-1 shows the marginal predictive densities at $h = 12$ for all five variables. The MC (solid black) and QMC (dashed blue) densities are kernel density estimates; Koopman (dash-dot green) is likewise a KDE. The Forward EP density (solid red) is the exact Gaussian mixture density (22).

Variable 1 displays the most pronounced bimodality, with a primary mode near -10 (bad-state region) and a secondary mode near 0 (good-state region), consistent with opposing intercepts and high persistence. All methods capture this bimodality; the EP mixture density displays it without the kernel smoothing that partially fills the inter-modal trough in the MC and QMC curves. Variables 2–5 exhibit moderate non-normality whose sign and magnitude are consistent with the skewness estimates in Table C-1. The close visual agreement between the EP curve and the MC benchmark across all five panels validates the EP approximation for this calibration.

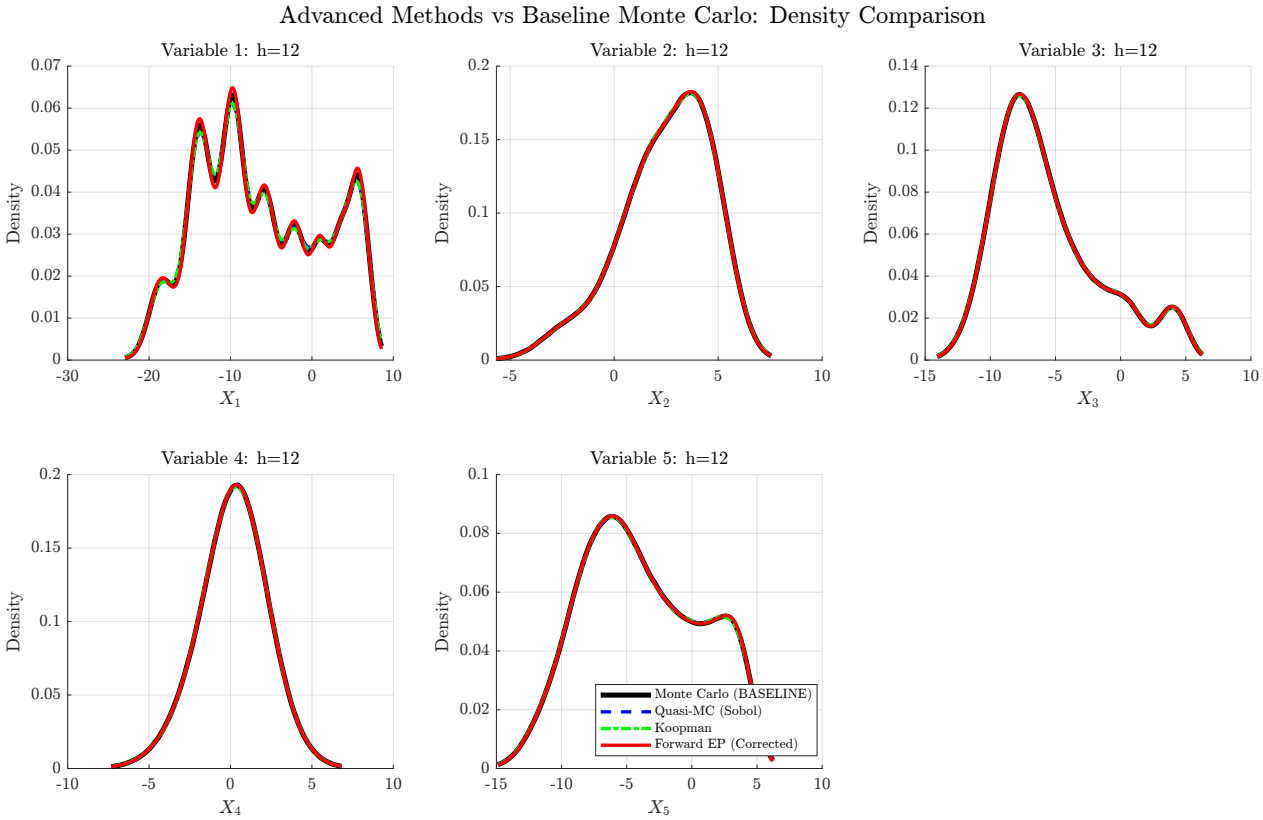


Figure C-1: Predictive densities at $h = 12$ for all five variables. Monte Carlo ($N = 10^7$, solid black) serves as the benchmark. Quasi-MC ($N = 10^6$, dashed blue) and Koopman ($N = 10^6$, dash-dot green) use kernel density estimation applied to their simulated draws. Forward EP (solid red) is the analytic Gaussian mixture density (22), evaluated without sampling. Variable 1 displays pronounced bimodality; variables 2–5 exhibit moderate non-normality consistent with Table C-1. The close alignment between the EP curve and the MC benchmark across all five panels validates the EP approximation for this calibration.

D Quantile IRF Methods: Technical Details

D.1 EP Forward-Recursion

For horizon H the EP filter enumerates all K^H regime paths $\kappa = (k_1, \ldots, k_H)$. Along each path the predictive distribution is propagated as follows.

Prediction step ($\ell = 1, \ldots, H$):

$$\mu_\ell^{\text{pred}} = c_{k_\ell} + B_{k_\ell}\, \mu_{\ell-1}, \tag{D-18}$$

$$\Sigma_\ell^{\text{pred}} = B_{k_\ell}\, \Sigma_{\ell-1}\, B_{k_\ell}' + Q_{k_\ell}. \tag{D-19}$$

EP update step (if $\ell < H$): because the next regime $k_{\ell+1}$ is observed and depends on $y_{t+\ell}$ through the probit link, the posterior for $y_{t+\ell}$ conditional on the next regime is a truncated Gaussian. Moment-matching yields:

$$\mu_\ell = \mu_\ell^{\text{pred}} + \text{sign}(k_{\ell+1})\, K_g\, r, \tag{D-20}$$

$$\Sigma_\ell = \Sigma_\ell^{\text{pred}} - \frac{r(a_{\text{std}} \cdot \text{sign}(k_{\ell+1}) + r)}{v} K_g K_g', \tag{D-21}$$

where $v = 1 + b_{k_\ell}' \Sigma_\ell^{\text{pred}} b_{k_\ell}$, $K_g = \Sigma_\ell^{\text{pred}} b_{k_\ell} / \sqrt{v}$, $a_{\text{std}} = (a_{k_\ell} + b_{k_\ell}' \mu_\ell^{\text{pred}}) / \sqrt{v}$, $r = \phi(a_{\text{std}}) / \Phi(\pm a_{\text{std}})$ is the inverse Mills ratio, and $\text{sign}(k_{\ell+1}) = +1$ if $k_{\ell+1} = 1$ and -1 otherwise.

Path weights incorporate the probit likelihood at each step; expected future transitions are integrated in closed form using $\mathbb{E}[\Phi(a + b'x)] = \Phi\Big((a + b'\mu)/\sqrt{1 + b'\Sigma b}\Big)$ for $x \sim \mathcal{N}(\mu, \Sigma)$. The full predictive density blends $2K^H$ Gaussian components over both initial regimes and all paths; quantiles are obtained by bisection on the mixture CDF.

D.2 Paired Monte Carlo Simulation

1. Allocate $N_k = \lfloor N\pi_k \rfloor$ simulations to initial regime $k \in \{0, 1\}$.

2. For each regime pool, run paired baseline–shocked simulations sharing the same sequence of uniform regime draws $\{u_i\}$ and standardized innovations $\{\varepsilon_i^{\text{std}}\}$. At each step h:

$$s_{h,i}^j = \mathbf{1}\left\{u_i < \Phi\left(a_{s_{h-1,i}^j} + b'_{s_{h-1,i}^j} Y_{h-1,i}^j\right)\right\},$$
$$Y_{h,i}^j = c_{s_{h,i}^j} + B_{s_{h,i}^j} Y_{h-1,i}^j + L_{s_{h,i}^j} \varepsilon_i^{\text{std}},$$

   where $Q_k = L_k L_k'$ and $j \in \{\text{base}, \text{shock}\}$ shares the same u_i and $\varepsilon_i^{\text{std}}$.

3. Pool draws across initial regimes and compute the QIRF by quantile matching at each horizon h:

$$\widehat{\text{QIRF}}_{\text{MC}}(\tau, h) = \hat{Q}_\tau\left(\{Y_{h,i}^{\text{shock}}\}\right) - \hat{Q}_\tau\left(\{Y_{h,i}^{\text{base}}\}\right). \tag{D-22}$$

   By the Glivenko–Cantelli theorem this converges almost surely to $\text{QIRF}(\tau, h)$ as $N \to \infty$.

Remark 4. *An alternative MC IRF takes the τ-quantile of the pairwise differences $Y_{h,i}^{shock} - Y_{h,i}^{base}$. Since $Q_\tau(X - Y) \neq Q_\tau(X) - Q_\tau(Y)$ for nonlinearly related random variables, this does not equal (24). When both scenarios share the same regime realization—which occurs with high probability at short horizons—the pairwise difference is nearly deterministic conditional on the regime, collapsing the quantile bands to discrete values and suppressing the full distributional heterogeneity captured by quantile matching.*

D.3 Validation

Figures D-2–D-3 document that the two methods produce virtually indistinguishable results across all horizons and quantiles, providing reassurance that neither the Gaussian-path

approximation of the EP filter nor the finite-sample error of the MC simulations ($N = 10^6$ draws) materially distorts the reported distributions.

Figure D-2 plots the quantile fan charts for the output gap and the stress indicator following a shock of +1.5 to the stress indicator, with EP-based fans on the left and MC-based fans on the right. The median trajectory (yellow), the "good-outcome" bands (blue), and the "bad-outcome" bands (red) align closely across methods at all horizons $h = 1, \ldots, 12$. Figure D-4 examines the joint predictive density of both variables at selected horizons; the EP analytical contours closely track the kernel density of the MC scatter, confirming that the Gaussian mixture captures the bivariate shape of the predictive distribution, including the off-diagonal co-movement induced by regime transitions. Figure D-3 overlays the MC histogram with the EP analytical density curve for each variable separately, showing that the left-skewed shape of the output-gap distribution at longer horizons—reflecting the larger variance and more adverse mean in the bad regime—is reproduced faithfully by both methods.

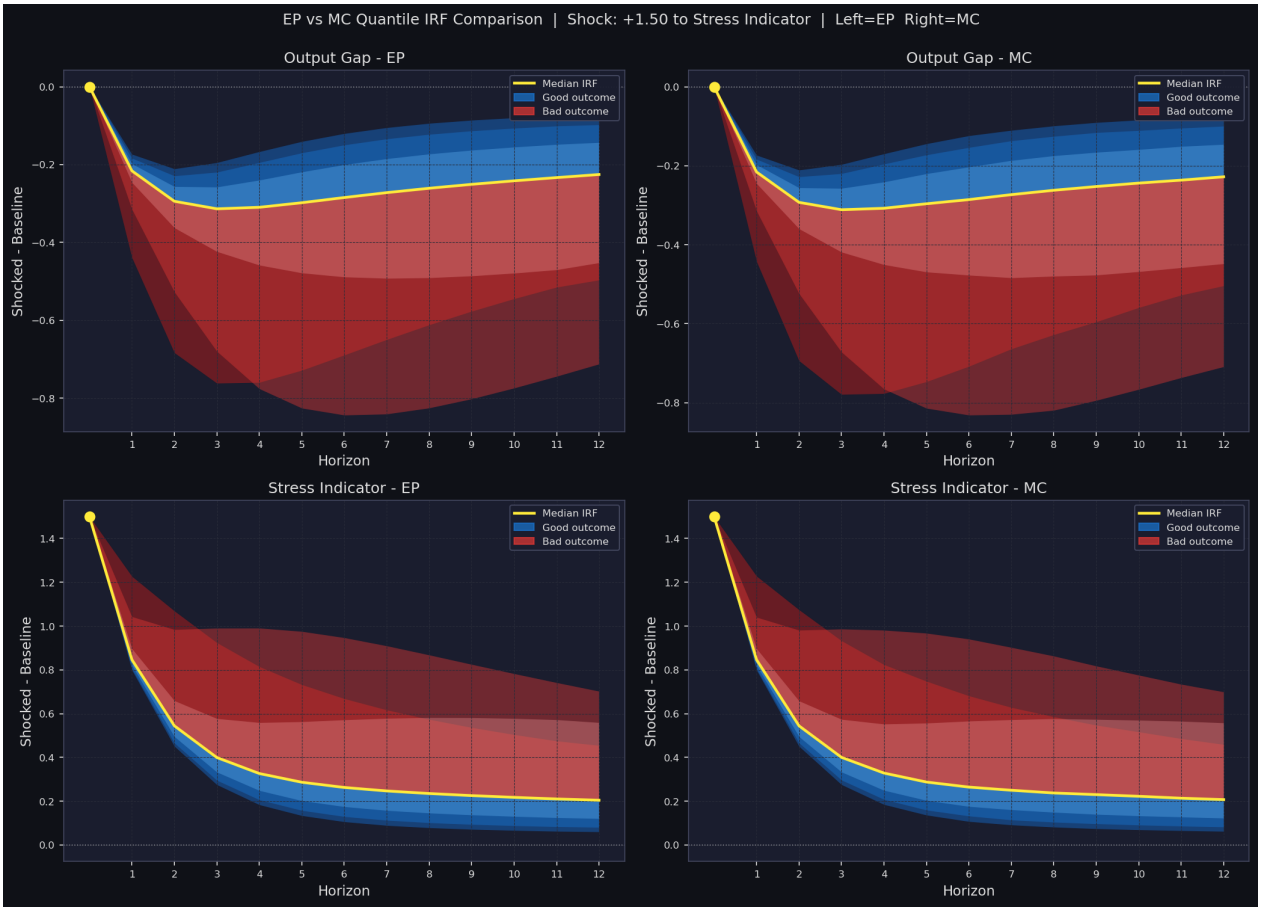


Figure D-2: Quantile IRF fan charts: EP (left column) vs. Monte Carlo (right column) for the output gap (top row) and stress indicator (bottom row). Shock: +1.5 to the stress indicator. Yellow line: median IRF. Blue (red) bands: upper (lower) quantile intervals corresponding to favorable (adverse) outcomes. Bands show the 10–90, 20–80, and 35–65 percentile ranges relative to the median.

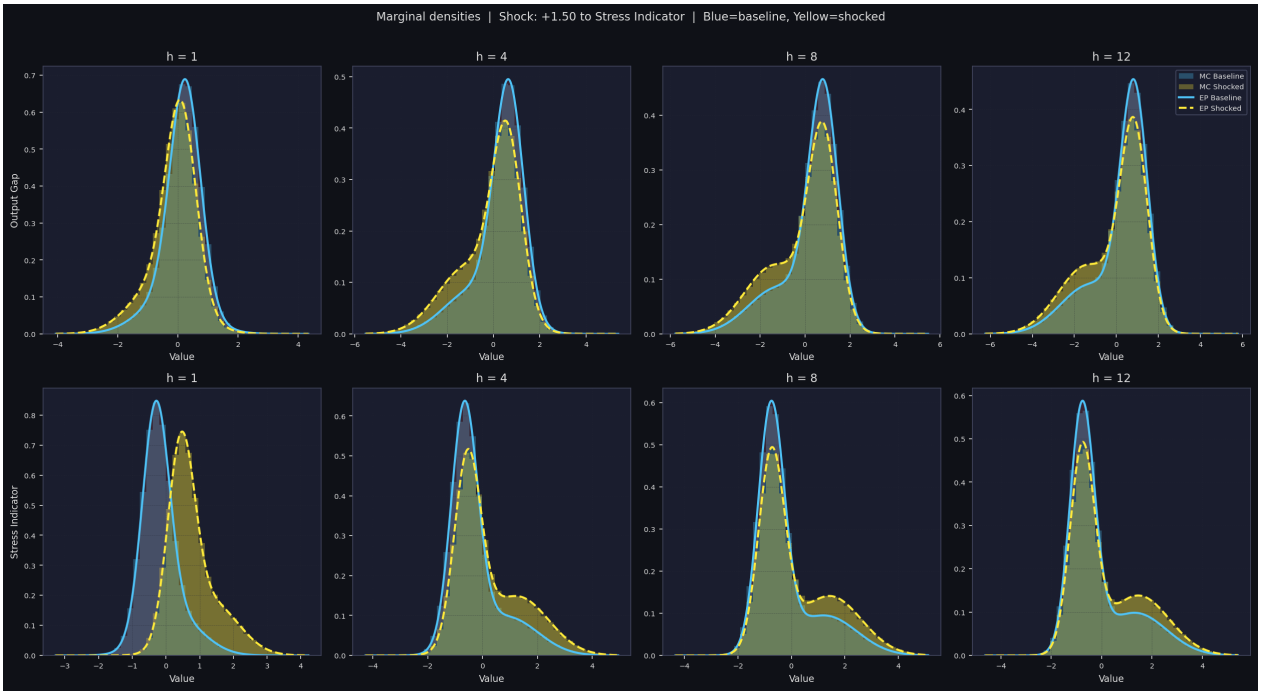


Figure D-3: Marginal predictive densities for the output gap (top row) and stress indicator (bottom row) at horizons $h \in \{1, 4, 8, 12\}$. Shaded areas: MC densities. Solid/dashed curves: EP analytical densities. Blue (yellow): baseline (shocked) scenario.

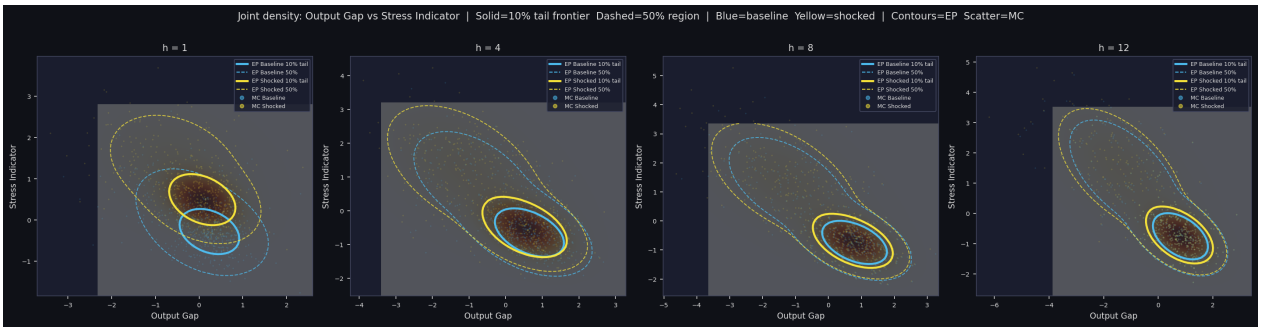


Figure D-4: Joint predictive density of the output gap and stress indicator at horizons $h \in \{1, 4, 8, 12\}$. Solid (dashed) contours show the EP 10% tail frontier (50% region) under the baseline (blue) and shocked (yellow) scenarios. Scatter points are MC draws.