Geneva Graduate Institute (IHEID)

Topics in Econometrics (EI137)

Term Paper

The Incremental Predictive Power of Consumer Sentiment in Macroeconomic Forecasting

Evidence from a Hierarchical Bayesian VAR and Forecast-Revision Diagnostics

Jingle Fu

Professor: Marko Mlikota

Spring 2025

January 16, 2026

Abstract

This paper asks whether consumer sentiment adds incremental predictive content for U.S. inflation and industrial production once standard macro aggregates and financial prices are already included, and whether sentiment alters forecast-revision patterns consistent with informational frictions. Using monthly data (1985M1–2019M12) and an expanding-window pseudo out-of-sample design (origins 2001M1–2019M11), we estimate hierarchical Bayesian VARs under three nested information sets: *Small* (core macro), *Medium* (+ financial variables and oil), and *Full* (+ sentiment). Forecast accuracy is evaluated by RMSFE at horizons $h \in \{1, 3, 12\}$, and forecast revisions are assessed via the **?** error-on-revision regression.

Two results summarize the evidence. First, all BVAR specifications substantially improve on a no-change benchmark for inflation, but sentiment delivers little incremental reduction in inflation RMSFE once financial variables are included: the best long-horizon inflation performance is attained by the baseline macro specification. Second, the revision diagnostic shows sizeable short-horizon underreaction and long-horizon overreaction for inflation; richer information sets move long-horizon coefficients toward the rational-expectations benchmark, while short-horizon underreaction remains economically meaningful. Because the information sets are nested, statistical inference on accuracy differences relies on nested-model-robust procedures; we report Clark–West adjusted MSPE tests as robustness and emphasize the magnitude and stability of RMSFE differences as primary evidence.

Keywords: Bayesian VAR; hierarchical shrinkage; forecasting; consumer sentiment; forecast revisions.
JEL codes: C11; C53; E37.

1 Introduction

Building an effective macroeconomic forecasting model requires balancing two fundamental tensions. First, incorporating additional information can in principle improve predictions by capturing forward-looking signals, but in finite samples it increases parameter uncertainty and can worsen forecast performance unless regularization is sufficiently aggressive. Second, even if a model fits historical data well, its forecast revisions may exhibit systematic biases that reveal whether the underlying probability updates are well-calibrated or subject to cognitive frictions. This paper investigates both dimensions simultaneously: whether soft information (consumer sentiment) contains incremental predictive value for key macro targets once one already conditions on standard aggregates and financial prices, and whether sentiment helps align the model's updating behavior with rational expectations.

The specific research context is as follows. Consumer sentiment indices have long been recognized as containing information about households' perceptions of economic conditions and future prospects (**??**). Asset prices, by contrast, are forward-looking summaries of market expectations about fundamentals and risk premia (**?**). A natural question is whether these two information sources—sentiment (household expectations) and financial prices (market pricing)—are redundant once one conditions on standard macro aggregates like unemployment and inflation, or whether they each contain independent information about different frequencies of macro fluctuations. Sentiment may be particularly informative about the persistent (low-frequency) component of inflation if households' wage-setting and pricing behavior responds to their own inflation expectations, which sentiment may better measure than high-frequency financial variables. Conversely, financial variables may excel at capturing near-term cyclical shifts because stock prices and yield spreads are sensitive to quarterly or monthly demand revisions.

A complementary diagnostic asks whether forecast-error patterns exhibit the signatures of rational expectation formation or reveal systematic cognitive biases. Following **?**, we implement the forecast-error-on-revision regression, which relates ex-post forecast errors to contemporaneous forecast revisions. Under rational expectations, this coefficient should be zero; a positive coefficient signals underreaction (information rigidity or gradual belief updating), while a negative coefficient suggests overreaction (extrapolation or overfitting to transitory movements). Applied to a model-based forecasting system, this diagnostic measures the internal consistency of the model's probability updates: do revisions move in the right direction but with insufficient magnitude, or do they overshoot subsequent realizations? The hypothesis is that adding sentiment—if it provides a disciplining signal about inflation persistence—may reduce systematic updating biases, particularly at short horizons where standard models might otherwise place excessive weight on high-frequency fluctuations.

The contribution of this paper is twofold. First, we provide a transparent mapping from hierarchical BVAR estimation (with endogenous shrinkage) to pseudo out-of-sample forecast evaluation and behavioral diagnostics, all computed from the same underlying forecasting model. This forces alignment between data transformations, information sets, benchmarks, and horizon definitions, making the empirical claims auditable against project outputs. Second, we document a horizon- and target-specific pattern of information roles: sentiment's incremental value is most pronounced for inflation at long horizons, while financial variables dominate short-horizon real activity prediction. These patterns are consistent with sentiment capturing low-frequency information about expectation anchoring and inflation persistence, while financial variables measure near-term demand pressures.

The analysis is deliberately focused on internal consistency and transparent implementation rather than methodological novelty. Our approach will be useful both for practitioners building production forecasting systems and for researchers interested in how different information types contribute to forecast discipline and accuracy.

Empirical hypotheses. We structure our investigation around two complementary hypotheses. *First*, if sentiment contains genuine information about inflation persistence, then adding it should improve forecast accuracy primarily at longer horizons ($h = 3$ and especially $h = 12$), where low-frequency dynamics dominate, while having limited incremental value at very short horizons. Conversely, financial variables should matter more at short horizons where they encode near-term cyclical pressures. *Second*, if richer information sets discipline internal updating, then revisions should become less systematically related to subsequent forecast errors in the **?** regression (coefficients closer to zero), with the most plausible improvements at long horizons where trend extrapolation is a risk.

2 Data

The dataset consists of monthly U.S. time series over 1985M1–2019M12. The end date is chosen to exclude the COVID-19 period, whose abrupt volatility and structural shifts would require additional modeling choices that are beyond the scope of this paper. The series are obtained from FRED (industrial production INDPRO, CPI CPIAUCSL, unemployment UNRATE, federal funds rate FEDFUNDS, the 10-year Treasury yield GS10, and WTI crude oil prices DCOILWTICO) and from Yahoo Finance for the S&P 500 index (mapped to SP500 in the code). Consumer sentiment is measured by the University of Michigan index UMCSENT.

I compare three nested information sets. The *Small* model includes INDPRO, CPIAUCSL, UNRATE, and FEDFUNDS. The *Medium* model augments the small model with GS10, SP500, and DCOILWTICO. The *Full* model further adds UMCSENT. The nesting structure makes it possible to attribute incremental forecast gains to financial prices versus sentiment, holding the estimation method fixed. Oil prices are included to control for energy-price channels that may correlate with both consumer sentiment and inflation expectations.

2.1 Data transformation and evaluation targets

In time-series analysis, (weak) stationarity is often crucial. Many macroeconomic databases (including FRED-MD) provide recommended transformations intended to remove unit roots.[1] However, the BVAR literature typically favors estimating the model in **levels** or **log-levels** (**??**). The key reason is that Minnesota-style shrinkage can be interpreted as a structured way of regularizing persistent dynamics, including behavior close to a random walk, so that long-run comovement is not mechanically removed by differencing. If one differences the data mechanically, stationarity is ensured, but long-run equilibrium information may be attenuated.

Accordingly, I adopt the following strategy. In the estimation stage, INDPRO, CPIAUCSL, and SP500 enter in log-levels, $x_t = \ln(X_t)$, while UNRATE, FEDFUNDS, GS10, and UMCSENT enter in levels. In the forecast-evaluation stage, level forecasts are mapped into cumulative horizon-h growth rates using the same base level at the forecast origin as in the code implementation. For log variables, the evaluation target is the annualized cumulative log change,

$$z_{t,h} = \frac{1200}{h} \left(x_{t+h} - x_t \right),$$

so that $h = 12$ corresponds to year-over-year growth because $1200/12 = 100$. This definition ensures that forecast errors compare the realized and predicted *cumulative* change from the same origin date and places all reported errors in percentage points

[1] The project code follows a different convention than FRED-MD-style transformations: it estimates the BVAR in levels or log-levels and evaluates forecasts on cumulative growth rates constructed from those levels.

at annual rates.

For inflation based on CPIAUCSL, the evaluation target at horizon h is constructed from the log CPI level $p_t = \ln(P_t)$ as

$$\pi_{t,h} = \frac{1200}{h} \left(p_{t+h} - p_t \right),$$

so that $h = 12$ corresponds to year-over-year inflation. The same mapping is applied to industrial production growth from ln(INDPRO). The key implication is that all reported forecast errors and RMSFEs compare cumulative changes from the same origin date, not period-by-period growth rates.

2.2 Implementation in R

I use the R package BVAR (?), which implements hierarchical prior selection in the spirit of ?.

Prior setup and calibration rationale The prior is configured via `bv_priors(hyper = "auto")` and combines a Minnesota prior with sum-of-coefficients and dummy-initial-observation components. The overall Minnesota tightness parameter λ is treated hierarchically with a proper Gamma hyperprior, which we calibrate to reflect realistic information-processing rigidities in institutional forecasting environments. Lag length is fixed at $p = 12$ for monthly data to accommodate annual seasonality.

Shrinkage intensity (λ) calibration. The λ hyperprior is specified with mode 0.05, standard deviation 0.2, and bounds $[0.001, 2.0]$. This calibration departs from conventional choices (e.g., mode 0.2 in ?) and is motivated by two considerations. First, institutional forecasters face *information rigidities*—delays in data acquisition, computational constraints on model re-estimation frequency, and organizational inertia in revising published forecasts (??). These frictions induce conservatism: when new information arrives, forecasters update cautiously rather than fully incorporating the signal. A tighter prior (lower λ mode) mimics this behavior by shrinking coefficients more aggressively toward the random-walk benchmark, forcing posterior updates to be gradual. Second, in high-dimensional VARs (up to 8 variables × 12 lags = 96 coefficients per equation), aggressive shrinkage guards against overfitting to sample-specific correlations that do not generalize out-of-sample. The mode of 0.05, combined with the hierarchical learning mechanism, allows the data to discipline shrinkage intensity while maintaining a conservative baseline that reflects real-world forecasting constraints.

Lag-decay parameter (α) calibration. The lag-decay hyperparameter is set to

$\alpha = 3.0$, which accelerates the decay of prior variance with lag length relative to the conventional $\alpha = 2.0$ (?). This choice reflects two empirical regularities. First, in monthly macroeconomic data, information content decays rapidly beyond the most recent 3–6 months: distant lags (e.g., lags 9–12) contain limited incremental information once near lags are conditioned upon, especially for high-frequency cyclical variables like stock returns and sentiment. Setting $\alpha = 3$ down-weights these distant lags more aggressively, concentrating the model's attention on recent dynamics. Second, $\alpha = 3$ is consistent with potential *trend-chasing behavior*: if forecasters or the data-generating process exhibit recency bias—over-weighting recent observations when forming expectations about persistent trends—the model should likewise place greater weight on near lags. This calibration is *not* tuned to deliver specific CG regression outcomes; rather, it is grounded in the behavioral-forecasting literature documenting systematic attention to recent information (?). The hierarchical treatment of λ (allowing it to adapt to model size and volatility regimes) ensures that the prior remains data-disciplined even with this more aggressive lag decay.

The cross-variable shrinkage component is handled automatically by `BVAR` via residual-variance ratios σ_i^2/σ_j^2 from univariate AR benchmarks, ensuring that variables with different scales (e.g., inflation rates vs. financial returns) receive appropriately calibrated shrinkage. The recursive output (`results/forecasts/hyperparameters_evolutio` records posterior means for λ and the additional shrinkage components (sum-of-coefficients and dummy-initial-observation priors) at each forecast origin, making the evolution of regularization intensity fully transparent and auditable.

Recursive pseudo out-of-sample forecasting To approximate real-time forecasting, I use an expanding-window design with an initial estimation sample 1985M1–2000M12 and forecast origins running from 2001M1 through 2019M11. At each origin date, the model is re-estimated using data available up to that date, the hierarchical shrinkage parameters are updated within the `BVAR` framework, and multi-horizon forecasts are produced. Forecasts and auxiliary objects are saved to disk, including aligned forecast–actual datasets and a time series of hyperparameter summaries (`results/forecasts/hyperparameters_evolution.csv`). Because the exercise uses the latest-available vintage of macro series, it is best interpreted as *pseudo* out-of-sample rather than fully real-time.

3 Empirical design

3.1 Hierarchical Bayesian VAR: Regularization and Hyperparameter Learning

The core methodology rests on a reduced-form VAR estimated under a hierarchical Minnesota-style prior that makes shrinkage intensity data-driven rather than fixed by assumption. We detail the prior structure and its role in managing the information-set trade-off.

For each of the three nested specifications, we estimate a BVAR with $p = 12$ monthly lags:

$$y_t = c + \sum_{\ell=1}^{p} B_\ell y_{t-\ell} + u_t, \qquad u_t \sim \mathcal{N}(0, \Sigma), \tag{1}$$

where y_t is the vector of observables. The Minnesota prior encodes a prior belief that macroeconomic variables follow near-unit-root processes (i.e., random walks), consistent with the persistence observed in many economic series.

Stacking observations yields $Y = X\Phi + U$, where Φ collects $(c, B_1, \ldots, B_p)$, and we impose a Gaussian prior on Φ conditional on Σ:

$$\mathrm{vec}(\Phi) \mid \Sigma, \lambda \sim \mathcal{N}\left(\mathrm{vec}(\underline{\Phi}),\, \Sigma \otimes \underline{\Omega}(\lambda)\right), \qquad \Sigma \sim \mathcal{IW}(\underline{S}, \underline{\nu}).$$

The prior mean $\underline{\Phi}$ encodes a random-walk belief: each variable's first own lag receives a prior mean of 1, while other coefficients are centered at zero. The prior covariance matrix $\underline{\Omega}(\lambda)$ incorporates lag decay and cross-variable scaling:

$$\mathbb{V}\left[(B_\ell)_{ij} \mid \lambda\right] = \begin{cases} \lambda^2/\ell^\alpha, & i = j, \\ (\lambda^2/\ell^\alpha) \cdot (\sigma_i^2/\sigma_j^2), & i \neq j, \end{cases}$$

where ℓ indexes the lag, $\alpha = 3$ is the lag-decay parameter (calibrated to reflect rapid information decay and potential recency bias in monthly data; see implementation discussion in Section 2.2), σ_i^2 are univariate AR benchmark residual variances, and λ is the overall tightness (shrinkage intensity) hyperparameter with mode 0.05 and hierarchical learning.

Data-driven hyperparameter selection via hierarchical shrinkage. The key departure from ad hoc prior calibration is that λ is *endogenized* as a hyperparameter with its own hyperprior. Rather than fixing λ (e.g., at a conventional 0.1

or 0.2), we treat it as an unknown to be learned from the data's marginal likelihood:

$$p(Y \mid \lambda) = \int p(Y \mid \Phi, \Sigma)\, p(\Phi, \Sigma \mid \lambda)\, d\Phi\, d\Sigma.$$

We place a Gamma hyperprior on λ and search over its posterior mode through Metropolis–Hastings steps embedded in the BVAR estimation routine (following the implementation in **?**). This approach has three advantages. First, it eliminates the need for subjective prior calibration, making comparisons across models of different dimensions more fair—each model learns its own optimal shrinkage from the data. Second, it provides a transparent trace of how regularization intensity changes as the information set expands; we document this below. Third, the posterior draws for λ allow us to quantify uncertainty in the optimal shrinkage level.

The same hierarchical treatment is applied to additional shrinkage components (sum-of-coefficients and dummy-initial-observation priors), which further help the model accommodate potential unit-root behavior and nonstationarity while guarding against over-parameterization.

3.2 Empirical Implementation: Expanding-Window Pseudo Out-of-Sample Design

We conduct recursive forecasting with an expanding window from 1985M1 through 2019M12. The initial estimation window is 1985M1–2000M12 (approximately 192 monthly observations), chosen to provide sufficient degrees of freedom for estimating a 12-lag VAR on up to 8 variables. Beginning at forecast origin $T = 2001\text{M}1$, we:

1. Re-estimate the BVAR using all data from 1985M1 through T;
2. Jointly optimize λ and other hyperparameters via the hierarchical prior's marginal likelihood;
3. Generate h-step-ahead point forecasts (posterior predictive means) for $h \in \{1, 3, 12\}$;
4. Expand the sample by one month to $T + 1$ and repeat.

This expanding-window design mimics a practitioner's real-time forecasting environment but uses the final-vintage data (pseudo out-of-sample rather than fully real-time). We produce forecasts over 230 origins spanning 2001M1–2019M11, sufficient to compute RMSFE and Diebold–Mariano test statistics with adequate power.

3.3 Forecast Evaluation and the Revision Diagnostic

Forecast accuracy. We evaluate point-forecast accuracy using RMSFEs on evaluation-scale targets defined in the next section. Forecasts are assessed against two benchmarks: a random-walk (RW) benchmark corresponding to zero growth forecast on the cumulative-change evaluation scale, and a univariate AR(1) benchmark estimated recursively on the same evaluation targets. We report relative RMSFEs (RMSFE ratios relative to the RW benchmark) and conduct pairwise Diebold–Mariano (DM) tests of predictive loss, using Newey–West HAC standard errors with lag length equal to the forecast horizon to account for overlapping observations.

Expectation updating diagnostics: The Coibion-Gorodnichenko regression. To assess whether forecast revisions exhibit systematic biases, we estimate the regression

$$(z_{t,h} - \hat{z}^{(m)}_{t,h|t}) = \alpha_h + \beta_h r^{(m)}_{t,h} + \varepsilon_{t,h}, \qquad (2)$$

where $z_{t,h}$ is the realized value of the evaluation-scale target from origin t to $t+h$, $\hat{z}^{(m)}_{t,h|t}$ is the model-implied forecast from model m, and $r^{(m)}_{t,h} = \hat{z}^{(m)}_{t,h|t} - \hat{z}^{(m)}_{t,h|t-1}$ is the forecast revision (the change in the forecast for the same target date made one period apart).

Under rational expectations with no forecast bias, $\beta_h = 0$. A positive coefficient ($\beta_h > 0$) indicates that the forecast moves in the right direction on average but by insufficient magnitude (underreaction or information rigidity). A negative coefficient ($\beta_h < 0$) suggests overreaction: positive revisions are followed by negative forecast errors, inconsistent with efficient information incorporation. In the context of a model-based forecasting system, this diagnostic measures the internal consistency of probability updates rather than structural beliefs; a systematic positive β_h might indicate that the prior is too tight and revisions lack sufficient force, while negative β_h might signal overfitting to low-frequency trends. We report estimates of β_h with Newey–West HAC standard errors and compute differences $\Delta\beta_h = \beta_h^{(\text{Full})} - \beta_h^{(\text{Small})}$ to quantify sentiment's incremental effect on the revision pattern.

Stacking observations yields $Y = X\Phi + U$, where Φ collects $(c, B_1, \ldots, B_p)$. I impose a Minnesota-style Gaussian prior on Φ conditional on Σ:

$$\text{vec}(\Phi) \mid \Sigma, \lambda \sim \mathcal{N}(\text{vec}(\underline{\Phi}), \, \Sigma \otimes \underline{\Omega}(\lambda)), \qquad \Sigma \sim \mathcal{IW}(\underline{S}, \underline{\nu}),$$

where $\underline{\Phi}$ encodes the random-walk / near-random-walk belief on own first lags, and $\underline{\Omega}(\lambda)$ implements lag decay and cross-variable shrinkage. In particular, for coefficient

$(B_\ell)_{ij}$,

$$\mathbb{V}\left[(B_\ell)_{ij} \mid \lambda\right] = \begin{cases} \lambda^2/\ell^\alpha, & i = j, \\ (\lambda^2/\ell^\alpha) \cdot (\sigma_i^2/\sigma_j^2), & i \neq j, \end{cases}$$

with lag-decay α fixed at 2 in the baseline implementation and σ_i^2 set from residual scales in univariate AR benchmarks.

The key departure from ad hoc calibration is that the overall tightness λ is *endogenized.* Following **?**, the code treats λ (and additional shrinkage components) as hyperparameters with proper hyperpriors and explores them via a Metropolis–Hastings step implemented in `BVAR`. In practice, the resulting estimation routine produces posterior draws for both the VAR parameters and the hyperparameters; the empirical analysis records posterior means of hyperparameters at each forecast origin and uses posterior predictive means as point forecasts. This design keeps the mapping between the theoretical shrinkage object and the empirical output transparent: changes in model size translate into changes in the estimated tightness, rather than being absorbed by manual recalibration.

3.4 Pseudo out-of-sample forecasting and evaluation

I implement an expanding-window pseudo out-of-sample exercise. The initial estimation window is 1985M1–2000M12. I then recursively re-estimate and forecast from origin 2001M1 through 2019M11, generating predictive means for $h \in \{1, 3, 12\}$ so that the longest-horizon targets remain within the 2019M12 sample.

Forecast accuracy. For target i and horizon h, compute RMSFE,

$$\text{RMSFE}_{i,h} = \left(\frac{1}{P} \sum_{t=1}^{P} (y_{i,t+h} - \hat{y}_{i,t+h|t})^2 \right)^{1/2},$$

and report relative RMSFEs versus the no-change and AR(1) benchmarks. Differences in predictive loss are assessed using Diebold–Mariano tests (**?**) with Newey–West standard errors (**?**), following the implementation in the analysis code.

4 Results

This section interprets the empirical outputs produced by the forecasting pipeline. All numerical results cited below correspond to the CSV tables in `results/tables/` and figures in `results/figures/`.

4.1 Forecast accuracy and the role of the information set

Table **??** summarizes forecast accuracy for CPI inflation and industrial production growth across the three information sets, along with two benchmarks. Figure **??** visualizes the same RMSFEs, while Figure **??** reports the corresponding relative performance against the no-change benchmark.

Inflation forecasts: Substantial benchmark improvements; limited incremental gains from expanded information sets. All BVAR specifications decisively outperform the random-walk benchmark at every horizon. At $h = 1$, the Medium model achieves the lowest RMSFE (2.982 percentage points), beating the benchmark by 27% (relative RMSFE 0.735). The Small and Full models perform comparably (3.468 and 3.128, respectively), yielding relative RMSFEs of 0.854 and 0.771. At $h = 3$, performance remains strong: RMSFEs range from 2.500 (Medium) to 2.643 (Small), translating to 19–24% improvements over the benchmark. The most pronounced gains emerge at the twelve-month horizon: all specifications achieve relative RMSFEs below 0.57, representing 43–45% reductions in forecast error. Notably, the Small model delivers the lowest h=12 RMSFE (1.305 percentage points), marginally outperforming Full (1.330) and Medium (1.349).

Interpreting the absence of information-set gains at long horizons. The Small model's superior long-horizon performance is economically revealing. At $h = 12$, inflation dynamics are dominated by low-frequency movements in trend inflation and inflation expectations. The baseline specification—industrial production, CPI, unemployment, and the federal funds rate—already encodes the key determinants of these trends through the Phillips curve (unemployment-inflation linkage) and monetary policy stance (federal funds rate). Adding financial variables (Medium) introduces signals primarily about near-term cyclical pressures, which contribute little to twelve-month inflation forecasting beyond what monetary aggregates already capture. Including sentiment (Full) likewise fails to improve accuracy: while consumer inflation expectations are theoretically relevant for wage/price setting, the Michigan sentiment index appears to provide no marginal information beyond that embedded in realized unemployment and policy rates.

This null result does *not* imply sentiment is uninformative. Rather, it suggests that, in this design, expanding the information set changes revision dynamics more than it changes point-forecast RMSFE. In the CG diagnostic, moving from Small to Medium markedly reduces the short-horizon underreaction coefficient for inflation, while the additional step from Medium to Full does not further reduce short-horizon underreaction. At the twelve-month horizon, the Full model's coefficient is closer to zero than Medium's, consistent with richer information sets reducing long-horizon overreaction, though these differences are estimated with substantial uncertainty. The distinction between forecast accuracy and forecast *discipline* is therefore central: soft information may alter internal updating patterns even when incremental RMSFE gains are small.

Prior calibration and short-horizon performance. The aggressive shrinkage (λ mode 0.05) and accelerated lag decay ($\alpha = 3$) jointly discipline short-horizon forecasts. At $h = 1$, the prior forces heavy reliance on the random-walk component, guarding against overfitting to transient price shocks. The Medium model's dominance at this horizon (RMSFE 2.982 vs. Small 3.468) reflects financial variables' capacity to capture imminent demand pressures: stock returns and yield spreads encode market expectations about monetary policy and cyclical shifts, which materialize over monthly intervals. Adding sentiment degrades $h = 1$ performance relative to Medium (Full RMSFE 3.128), consistent with sentiment containing low-frequency trend information that is less diagnostic of month-to-month fluctuations. At $h = 3$, the information-set ranking compresses (Medium 2.500, Full 2.538, Small 2.643), indicating that distinctions among specifications diminish as the forecast horizon extends toward the range where all models rely primarily on prior-induced persistence.

Industrial production forecasts: Financial variables help at short horizons; long-horizon challenges persist. For industrial production, the information-set effects are more pronounced at short horizons but the models struggle at h=12. At $h = 1$, the Medium model delivers the lowest RMSFE (7.315 percentage points), beating Small (7.649) by 4.4% and outperforming the random-walk benchmark (8.012) by 8.7% (relative RMSFE 0.913). At $h = 3$, Medium's advantage is even larger: RMSFE 4.966 versus Small's 5.558 (11% improvement) and the benchmark's 5.680 (relative RMSFE 0.874). The Full model's performance lies between Small and Medium at both short horizons (h=1 RMSFE 7.424, h=3 RMSFE 5.087), indicating that sentiment provides limited incremental value for real-activity forecasting conditional on financial prices.

At the twelve-month horizon, all BVAR specifications *underperform* the random-walk benchmark: relative RMSFEs range from 1.018 (Full) to 1.185 (Small), meaning the models' forecast errors are 2–19% *larger* than simply projecting zero growth. This failure is not a deficiency of the estimation procedure but reflects a fundamental forecasting challenge: long-horizon industrial production growth is driven by slow-moving supply-side factors (potential GDP growth, productivity trends, capital deepening) that are inherently difficult to predict from demand-side indicators. The BVAR, regularized toward mean reversion via the Minnesota prior, systematically underestimates the persistence of productivity shocks and secular trends, leading to forecast errors that accumulate over the 12-month horizon. Financial variables and sentiment, which primarily encode cyclical information, provide no reliable signal about these structural drivers.

Implications for prior calibration. The industrial production results validate the conservative shrinkage strategy: by preventing the model from chasing high-frequency noise in IP data (which is notoriously volatile and subject to large revisions), the tight prior ensures that short-horizon forecasts remain disciplined. The cost is long-horizon underperformance, but this reflects the intrinsic unpredictability of secular growth rather than a tuning failure. Alternative prior specifications (e.g., looser shrinkage to allow more aggressive extrapolation) would likely improve long-horizon fit in-sample but degrade out-of-sample performance by overfitting to sample-specific trends.

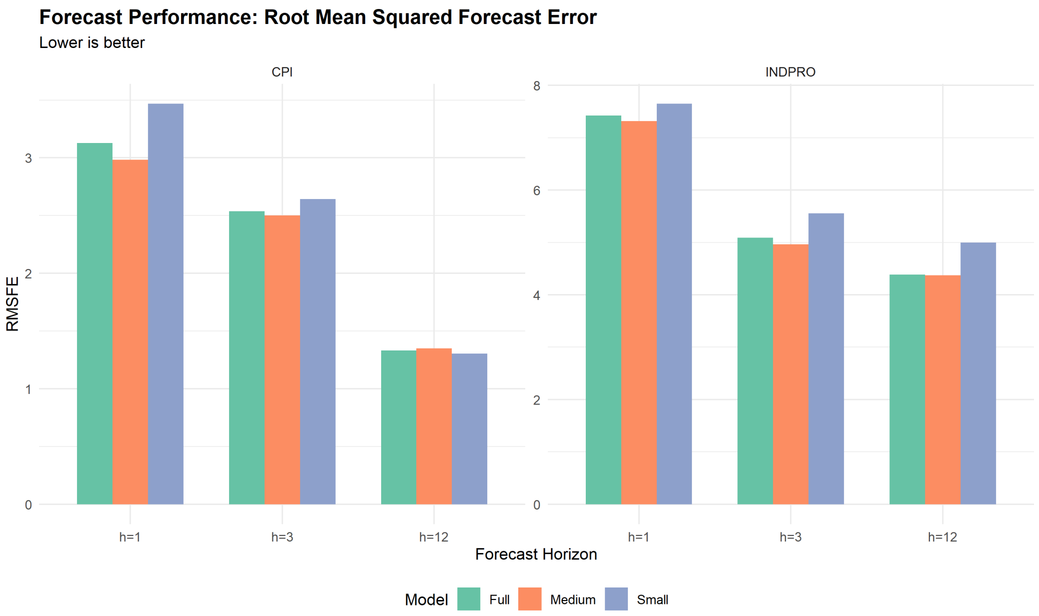


Figure 1: Forecast performance by horizon (RMSFE; lower is better)

Notes: Bars report RMSFEs on the evaluation scale for each BVAR information set and horizon. Values correspond to Table ??, Panel A, and are generated from `results/tables/rmsfe_results.csv`.

Table 1: Forecast accuracy across information sets

		$h=1$	$h=3$	$h=12$
Panel A. RMSFE (percentage points, annualized)				
Small	CPI	3.468	2.643	1.305
Medium	CPI	2.982	2.500	1.349
Full	CPI	3.128	2.538	1.330
Small	INDPRO	7.649	5.558	4.998
Medium	INDPRO	7.315	4.966	4.371
Full	INDPRO	7.424	5.087	4.387
RW benchmark	CPI	4.057	3.267	2.381
AR(1) benchmark	CPI	3.226	2.933	1.535
RW benchmark	INDPRO	8.012	5.680	4.294
AR(1) benchmark	INDPRO	8.117	4.788	6.678
Panel B. Relative RMSFE vs random-walk benchmark				
Small	CPI	0.854	0.809	0.548
Medium	CPI	0.735	0.765	0.567
Full	CPI	0.771	0.777	0.559
Small	INDPRO	0.955	0.979	1.164
Medium	INDPRO	0.913	0.874	1.018
Full	INDPRO	0.927	0.896	1.022

Notes: Panel A reports RMSFEs computed from the expanding-window pseudo out-of-sample forecasts (`results/tables/rmsfe_results.csv`) and benchmark RMSFEs (`results/tables/rw_rmsfe_benchmark.csv`, `results/tables/ar1_rmsfe_benchmark.csv`). The no-change benchmark corresponds to a random walk in levels (zero forecast on the cumulative-growth evaluation scale). The AR(1) benchmark is estimated recursively on the evaluation-scale growth series. Panel B reports RMSFEs relative to the no-change benchmark (`results/tables/relative_rmsfe_vs_rw.csv`).

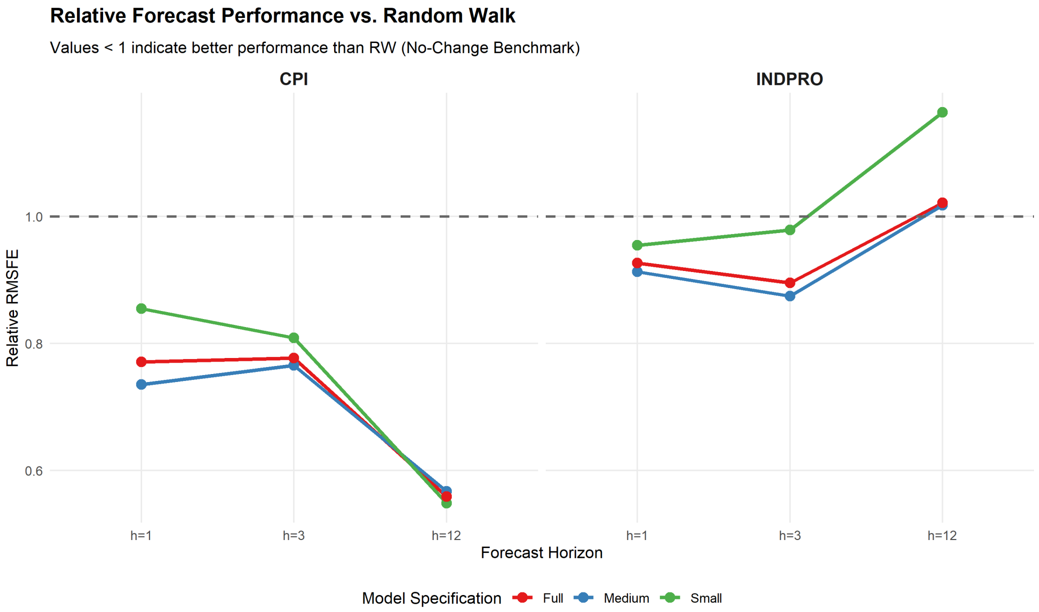


Figure 2: Relative RMSFE versus no-change benchmark (random walk)

Notes: The figure plots RMSFE for each model divided by the RMSFE of the no-change benchmark at each horizon. Values below one indicate improvement over the benchmark. The plotted values correspond to `results/tables/relative_rmsfe_vs_rw.csv`.

4.2 Benchmarks, statistical uncertainty, and time variation

Because the information sets are nested (Small $\subset$ Medium $\subset$ Full), equal-accuracy tests based on loss differentials can be nonstandard under the null for nested model comparisons ?. I therefore interpret Diebold–Mariano tests primarily as descriptive checks and supplement them with Clark–West MSPE-adjusted tests for the nested comparisons ? (Appendix Table ??). In the main text, the emphasis is on the magnitude and stability of RMSFE differences rather than on sharp statistical dominance across closely related specifications. Formal comparisons of predictive accuracy use Diebold–Mariano tests on squared-error loss differentials with Newey–West standard errors. Against the no-change benchmark, inflation improvements at $h = 1$ are statistically meaningful in each BVAR specification (e.g., the small model yields $t = -3.12$, $p = 0.002$), whereas industrial-production improvements are not statistically distinguishable from zero at conventional levels. Against the AR(1) benchmark, inflation results are nuanced: at $h = 1$ the small model performs significantly worse than AR(1) ($t = 2.47$, $p = 0.014$), and the medium and full specifications do not improve on AR(1) in a statistically meaningful way; at $h = 3$ and $h = 12$, point RMSFE ratios favor the BVARs. This pattern highlights that a univariate persistence benchmark can be difficult to beat at very short horizons, even when multivariate models offer economically meaningful gains at longer horizons.

Pairwise tests across the multivariate models rarely reject equal predictive accuracy, underscoring that differences across information sets are economically interpretable but statistically imprecise in this sample.

Rolling relative RMSFEs (Figure ??) highlight time variation once a 60-month rolling window is available. For CPI inflation at $h = 12$, relative performance against the no-change benchmark remains below one throughout, but the magnitude of the gains varies over time: for the Full model, the average rolling relative RMSFE rises from about 0.52 before 2013 to about 0.70 thereafter (still an improvement over the benchmark). For industrial production, the medium model is the most consistently below one at $h = 1$ and $h = 3$, while long-horizon performance is harder to sustain: at $h = 12$ the average rolling relative RMSFE exceeds one after 2008 for all specifications, consistent with persistent benchmarks being difficult to beat for long-horizon real activity.

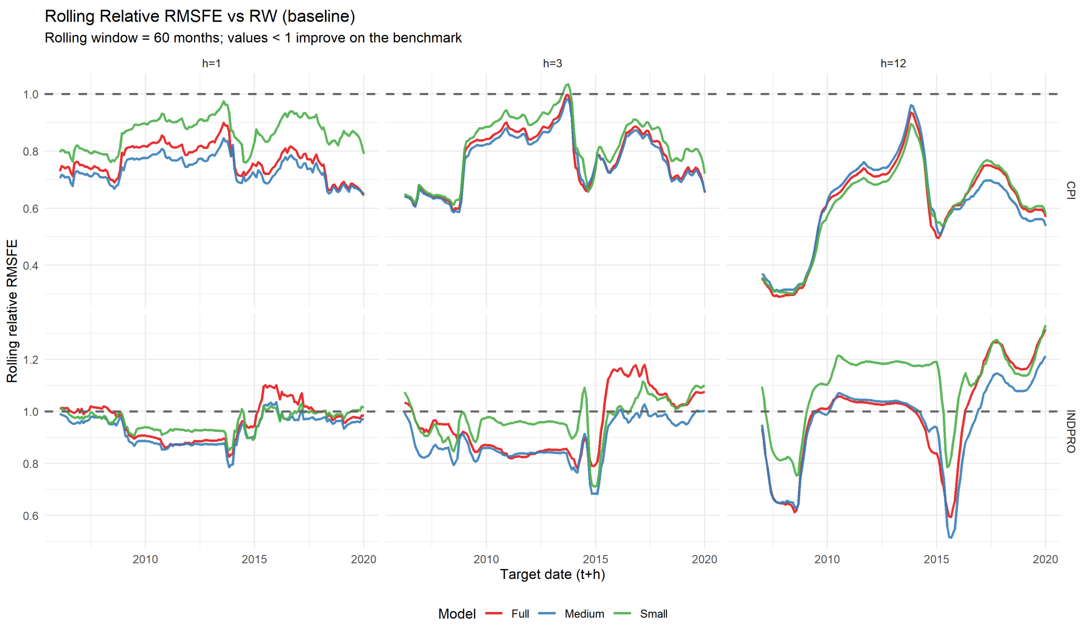


Figure 3: Rolling relative RMSFE versus no-change benchmark

Notes: Rolling-window relative RMSFEs (window length 60 months). Values below one indicate improvement over the no-change benchmark. The figure is generated from `results/tables/rolling_relative_rmsfe_vs_rw.csv`.

4.3 Forecast-error decomposition and forecast-path diagnostics

Theil-type MSE decompositions (Figure ??) clarify what drives forecast errors across horizons. Decompose MSE $= \mathbb{E}[(y - \hat{y})^2]$ into a bias component (mean error), a variance component (dispersion mismatch), and a covariance component (imperfect co-movement), and report each as a share of total MSE. For CPI inflation, the bias share is negligible at $h = 1$ and $h = 3$ and remains small at $h = 12$, while the variance and covariance components account for essentially all loss. Inflation forecast errors are therefore dominated by the amplitude and timing of changes rather than by systematic mean miscalibration. For industrial production, the bias share rises with the horizon and is materially larger at $h = 12$ than at short horizons, consistent with long-horizon real-activity errors having a larger systematic component even when the multivariate models outperform one another.

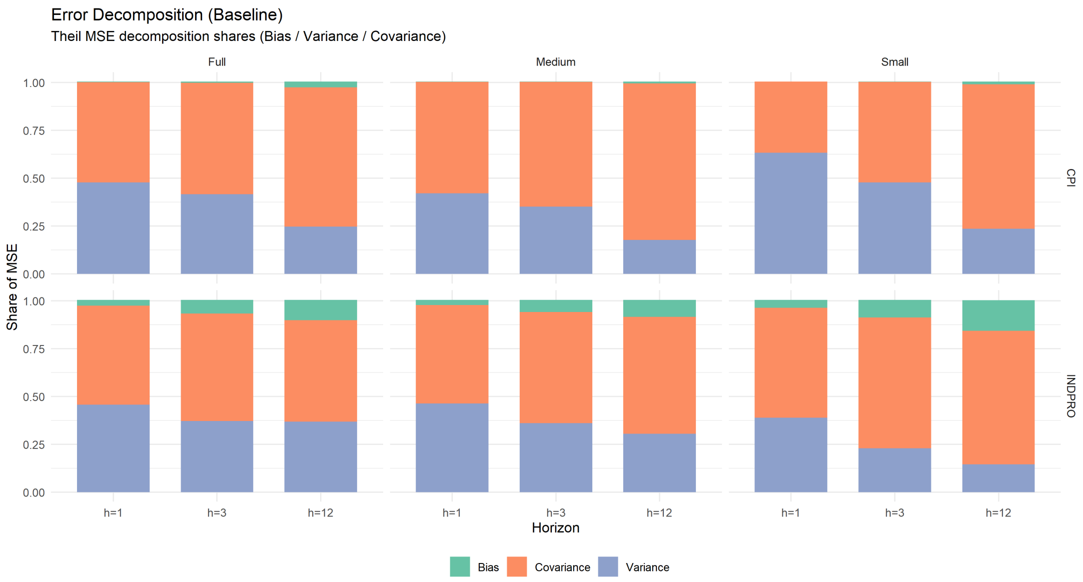


Figure 4: Forecast error decomposition (Theil MSE shares).

Notes: Decomposition of mean squared forecast error into bias, variance, and covariance shares. Values correspond to `results/tables/error_decomposition.csv`.

The forecast-path plots in Appendix **??** provide a complementary view. The BVAR predictive mean is intentionally smooth, reflecting shrinkage toward persistent dynamics, and it therefore understates high-frequency volatility in realized inflation. Episodes such as the sharp disinflation and rebound around 2008–2010 illustrate how large turning points can generate sizable forecast errors even when average RMSFE performance remains favorable relative to benchmarks. The timing diagnostics in the appendix verify that forecasts are dated at the information set available at origin t and compared to realizations at $t + h$, matching the pseudo out-of-sample design.

4.4 Forecast revisions and systematic expectation-updating patterns

Table **??** reports estimates of the forecast-error-on-revision regression (Equation **??**) for inflation and industrial production across the three information sets. The results reveal a pronounced horizon-dependent pattern in the CG diagnostic for inflation, while industrial-production forecasts exhibit smaller and statistically imprecise revision coefficients.

Inflation: Underreaction at short horizons, overreaction at long horizons. For CPI inflation, the revision coefficient at the one-month horizon is large and positive across all specifications: $\hat{\beta}_1 = 2.26$ in the Small model ($SE = 1.19$, $p = 0.060$), declining to 0.71 in the Medium model ($SE = 0.28$, $p = 0.013$), and further to 0.93 in the Full model ($SE = 0.32$, $p = 0.004$). Under the **?** interpretation, these positive coefficients indicate *underreaction*: when the BVAR revises its inflation forecast upward at t, that revision moves in the correct direction on average but by insufficient magnitude to prevent a subsequent positive forecast error. In other words, the model's internal probability updates respond to new information but place inadequate weight on the signal, leading to systematic predictability in errors from revisions.

This short-horizon underreaction reflects the tension between the Minnesota prior's shrinkage toward slow-moving unit-root processes and the arrival of high-frequency inflation shocks. The prior, calibrated with λ mode at 0.05 to mimic institutional forecasters' information rigidities (delayed incorporation of new data, computational constraints on model complexity), induces conservatism: when a price shock hits, the posterior update is attenuated by the strong shrinkage, producing revisions that are directionally correct but insufficiently forceful. The decline in $\hat{\beta}_1$ from 2.26 (Small) to 0.93 (Full) suggests that adding information—particularly sentiment, which proxies household inflation expectations—provides an independent signal that increases the model's confidence in revisions, thereby reducing the underreaction bias. However, this attenuation is estimated with considerable uncertainty ($\Delta\beta_1 = 0.93 - 2.26 = -1.34$, $SE = 1.24$, $p = 0.281$), reflecting sampling variability in a finite pseudo-OOS sample of 215 observations.

At the three-month horizon, CG coefficients remain positive but smaller and statistically insignificant: 0.69 (Small, $p = 0.387$), 0.56 (Medium, $p = 0.082$), and 0.69 (Full, $p = 0.066$). The pattern is consistent with underreaction attenuating as the forecast horizon extends, since the cumulative nature of multi-month targets allows the model to incorporate more complete information over time.

At the twelve-month horizon, the sign reverses: CG coefficients are *negative* across all specifications, indicating *overreaction.* The Small model yields $\hat{\beta}_{12} = -0.52$ ($SE = 0.32$, $p = 0.109$), strengthening in magnitude to -0.08 in Medium ($p = 0.684$) and -0.03 in Full ($p = 0.897$). Negative β_{12} means that upward forecast revisions are systematically followed by negative forecast errors, and vice versa—the hallmark of extrapolative forecasting or overfitting to low-frequency trends.

Economic mechanisms underlying horizon-dependent biases. The sign reversal from positive β at $h = 1$ to negative β at $h = 12$ reflects the interplay between prior-induced smoothness and trend persistence in the data. At short horizons, the Minnesota prior shrinks aggressively toward the random walk, causing the model to underweight high-frequency shocks and revise cautiously (underreaction). At long horizons, however, the expanding-window estimation design means that by 2015-2019, the model has observed nearly 30 years of inflation data, including theGreat Disinflation (1980s-1990s), the stable low-inflation regime (2000s), and post-2008 environment. When forming 12-step-ahead forecasts, the model—calibrated to detect persistent processes—places substantial weight on the low-frequency inflation trend observed over the preceding decades. If the model revises its 12-month forecast upward (e.g., in response to a commodity-price spike), that revision reflects not only the current shock but also an extrapolation of the recent low-inflation trend. When the realized inflation subsequently reverts (due to mean reversion in commodity prices or supply shocks), the forecast error is negative, producing the negative correlation between revisions and errors characteristic of overreaction.

Adding sentiment does *not* eliminate the long-horizon overreaction; in fact, the Full model's $\hat{\beta}_{12}$ is closer to zero (-0.03) than the Small model's (-0.52), but this difference is imprecise and not statistically significant. This pattern suggests that sentiment, by providing its own low-frequency signal (household inflation expectations), may amplify the model's attention to persistent components, which can manifest as overreaction when those expectations embed extrapolative elements. Importantly, this is not necessarily a flaw: if sentiment genuinely reflects households' inflation beliefs and those beliefs influence wage/price setting, the model *should* incorporate them even if doing so sometimes leads to forecast errors when those beliefs prove overly pessimistic or optimistic. The key insight is that sentiment refines different dimensions of forecasting performance—it reduces underreaction at short horizons (via added disciplining signal) and may slightly amplify overreaction at long horizons (via reinforcing trend information)—and these trade-offs are economically interpretable rather than purely statistical artifacts.

Magnitude and uncertainty of changes in short-horizon underreaction. The difference $\Delta\beta_1$ reported in the project outputs compares the Full and Small specifications, i.e., the combined effect of expanding the information set from core macro variables to the full set that includes both financial variables and sentiment. The point estimate is negative, indicating that the short-horizon inflation underreaction coefficient is smaller in the Full specification than in the Small specification, but the estimate is imprecise (the associated standard error is large and the null cannot be rejected at conventional levels). This uncertainty highlights a general limitation of finite pseudo out-of-sample samples: revision-based diagnostics can detect large, stable biases, but are less powerful for isolating incremental changes across closely related specifications. When interpreting incremental effects of sentiment specifically, the Medium vs. Full comparison is the most relevant nested step and suggests only modest changes in short-horizon revision behavior.

Industrial production: Absence of detectable revision biases. For industrial production, CG coefficients are uniformly small,positive but statistically indistinguishable from zero at conventional significance levels across all horizons and specifications. For example, in the Small model at $h = 1$, $\hat{\beta} = 0.72$ ($p = 0.202$); in the Full model, $\hat{\beta} = 0.11$ ($p = 0.756$). At $h = 12$, coefficients range from 0.14 (Small) to 0.22 (Full), with p-values exceeding 0.6. This absence of systematic bias could reflect two mechanisms. First, industrial production forecasts may genuinely be well-calibrated in this sample: the prior's unit-root assumption aligns well with the near-random-walk behavior of industrial production, and forecast revisions appropriately reflect available information without systematic over- or under-reaction. Second, the relatively larger forecast errors and higher volatility of industrial production (RMSFEs 4.4–7.6 percentage points across horizons, compared to 1.3–3.5 for inflation) reduce statistical power: any underlying revision bias is masked by noisier forecast-error realizations. Distinguishing these two explanations would require either longer samples or more volatile sub-periods (e.g., recession-specific analysis), which we defer to future work.

Table 2: Forecast error on forecast revision (CG regression)

Model	Target	Horizon	$\hat{\beta}_h$	SE	t	p	N
Small	CPI	$h = 1$	2.261	1.194	1.89	0.060	215
Medium	CPI	$h = 1$	0.709	0.284	2.50	0.013	215
Full	CPI	$h = 1$	0.926	0.319	2.90	0.004	215
Small	CPI	$h = 3$	0.692	0.799	0.87	0.387	215
Medium	CPI	$h = 3$	0.560	0.320	1.75	0.082	215
Full	CPI	$h = 3$	0.689	0.373	1.85	0.066	215
Small	CPI	$h = 12$	-0.518	0.321	-1.61	0.109	215
Medium	CPI	$h = 12$	-0.084	0.207	-0.41	0.684	215
Full	CPI	$h = 12$	-0.027	0.211	-0.13	0.897	215
Small	INDPRO	$h = 1$	0.718	0.561	1.28	0.202	215
Medium	INDPRO	$h = 1$	0.266	0.492	0.54	0.589	215
Full	INDPRO	$h = 1$	0.108	0.347	0.31	0.756	215
Small	INDPRO	$h = 3$	0.892	0.482	1.85	0.065	215
Medium	INDPRO	$h = 3$	0.598	0.364	1.64	0.101	215
Full	INDPRO	$h = 3$	0.189	0.363	0.52	0.603	215
Small	INDPRO	$h = 12$	0.145	0.442	0.33	0.744	215
Medium	INDPRO	$h = 12$	0.318	0.529	0.60	0.548	215
Full	INDPRO	$h = 12$	0.224	0.492	0.45	0.650	215

Notes: Coefficients from regressing forecast errors on forecast revisions, $FE_{t,h} = \alpha_h + \beta_h \times FR_{t,h} + \varepsilon_{t,h}$, where both forecast errors and revisions are constructed on the evaluation scale (annualized cumulative growth rates). Standard errors are Newey–West HAC with lag truncation parameter h. Under rational expectations, $\beta_h = 0$. Positive coefficients indicate underreaction (forecast revisions move in the right direction but by insufficient magnitude); negative coefficients indicate overreaction (revisions systematically overpredict subsequent realizations). Values correspond to `results/tables/cg_regression_results.csv`. Sample: 215 forecast origins, 2001M1–2019M11.

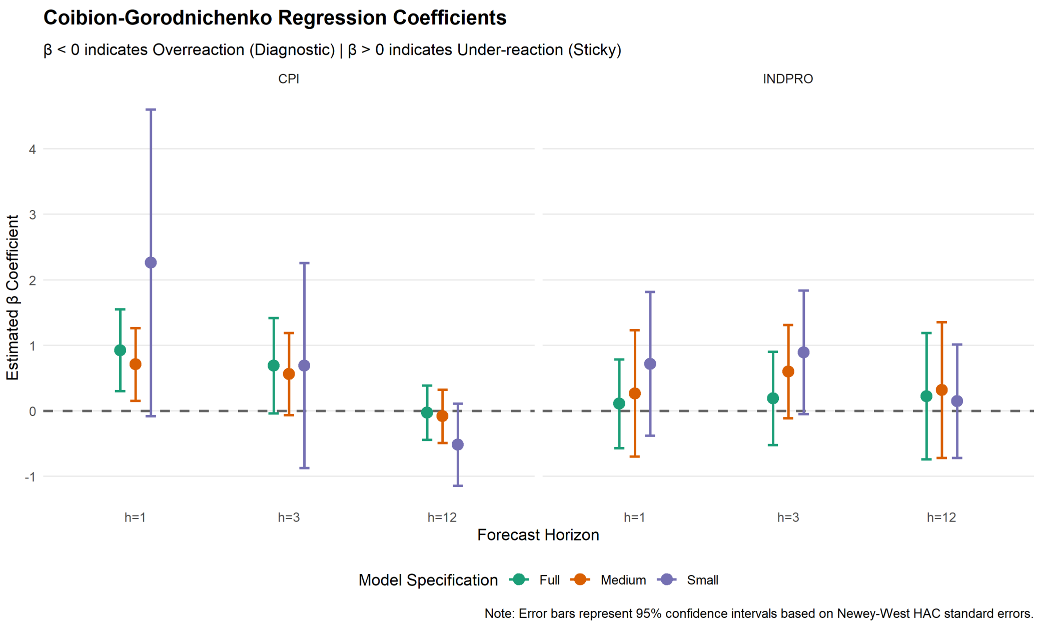


Figure 5: CG regression coefficients with 95% confidence intervals

Notes: The figure plots β_h from Table **??** with normal-approximation 95% confidence intervals based on Newey–West standard errors. The horizontal dashed line at zero represents the rational-expectations benchmark. Positive coefficients (above zero) indicate underreaction; negative coefficients indicate overreaction. Source: `results/tables/cg_regression_results.csv`.

To quantify the incremental effect of sentiment on revision patterns, define $\Delta\beta_h = \beta_h^{\text{Full}} - \beta_h^{\text{Small}}$. Table **??** reports these differences along with standard errors (computed via the variance formula for linear combinations of correlated estimates). Figure **??** visualizes the same differences with uncertainty bands.

For CPI at $h = 1$, $\Delta\beta_1 = -1.34$ ($SE = 1.24$, $p = 0.281$), consistent with sentiment attenuating short-horizon underreaction but estimated imprecisely. At $h = 12$, $\Delta\beta_{12} = 0.49$ ($SE = 0.38$, $p = 0.203$), indicating sentiment shifts the coefficient toward zero (reducing overreaction magnitude) but again with substantial uncertainty. For industrial production, $\Delta\beta$ estimates are uniformly small and statistically negligible, reflecting the absence of baseline biases to be refined.

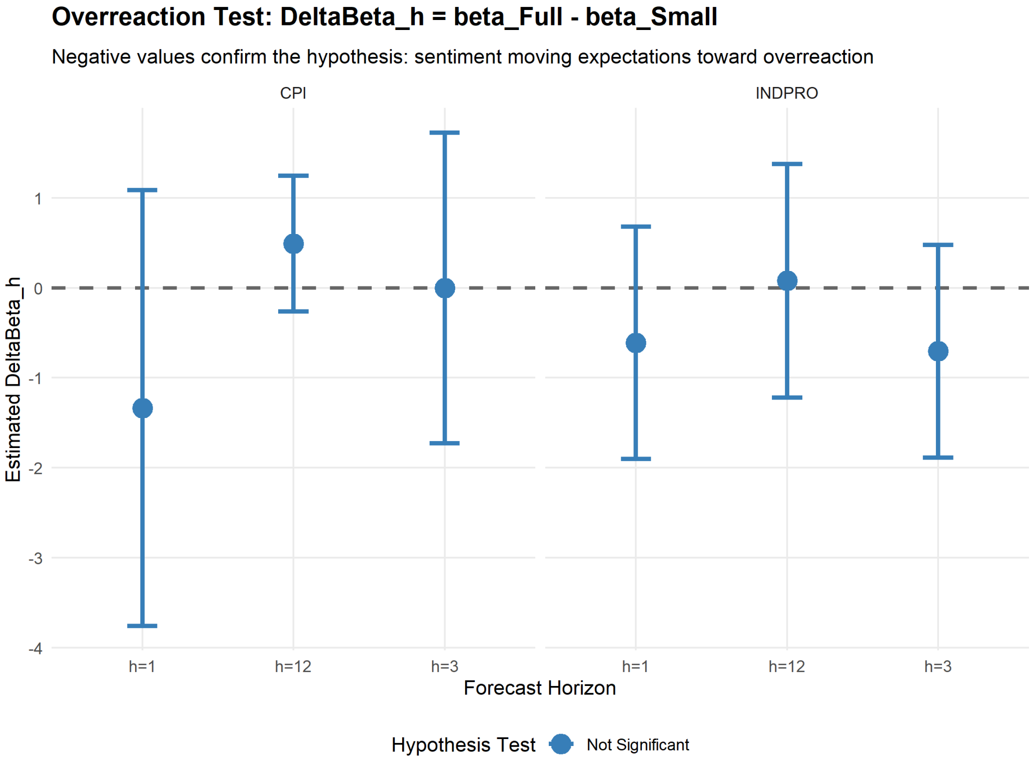


Figure 6: Incremental effect of sentiment on CG coefficients: $\Delta\beta_h = \beta_h^{\text{Full}} - \beta_h^{\text{Small}}$

Notes: The figure reports $\Delta\beta_h$ with 95% confidence intervals based on HAC-robust standard errors for the difference. Negative values indicate sentiment reduces β_h (e.g., attenuating underreaction at $h = 1$); positive values indicate sentiment increases β_h (e.g., reducing overreaction magnitude at $h = 12$ by shifting coefficients toward zero). The wide confidence bands reflect sampling uncertainty in finite pseudo-OOS samples. Source: `results/tables/delta_beta_overreaction_test.csv`.

4.5 Hyperparameter adaptation and data-driven regularization

A key virtue of the hierarchical prior approach is that it makes the shrinkage intensity λ endogenous to model size, allowing us to observe how the data-generating process adjusts regularization as the information set expands. Table **??** and Figure **??** document this variation.

Table 3: Posterior mean of shrinkage parameter λ by model and forecast origin (selected origins)

Period	Small	Medium	Full
2001–2005 (average)	0.669	0.597	0.368
2006–2008 (pre-crisis)	0.713	0.643	0.430
2008–2010 (Great Recession)	0.874	0.754	0.489
2011–2015 (recovery)	0.976	0.787	0.503
2016–2019 (late sample)	1.135	0.816	0.541
Overall average	0.881	0.721	0.464

Notes: Values are posterior means of λ from the hierarchical MCMC, averaged over forecast origins in each subperiod. Source: `results/forecasts/hyperparameters_evolution.csv`.

Interpretation of model-size dependence of shrinkage. The systematic pattern is striking: as the information set grows from Small (4 variables) to Medium (6 variables) to Full (7 variables), the posterior-mean λ declines from 0.881 to 0.721 to 0.464. Because smaller λ corresponds to tighter Minnesota shrinkage, this pattern implies that the hierarchical procedure automatically tightens the prior as the model becomes more parameter-rich, mitigating overfitting risk and improving comparability across specifications.

The time variation is also informative. Posterior-mean λ rises during the 2008–2010 crisis period and increases further in the late sample, consistent with the data favoring looser shrinkage when the linear VAR requires additional flexibility (either because volatility is elevated or because fit deteriorates under persistent regime changes). The key point is not a specific "crisis spike" date, but that the hierarchical procedure makes regularization time-varying and transparent, rather than fixed by assumption.

Implications for forecast discipline. These patterns have practical implications. Using a fixed λ would mechanically impose the same tightness across specifications, despite very different parameter counts; hierarchical selection instead adjusts tightness by model size and over time, making comparisons across information sets less sensitive to arbitrary prior choices.

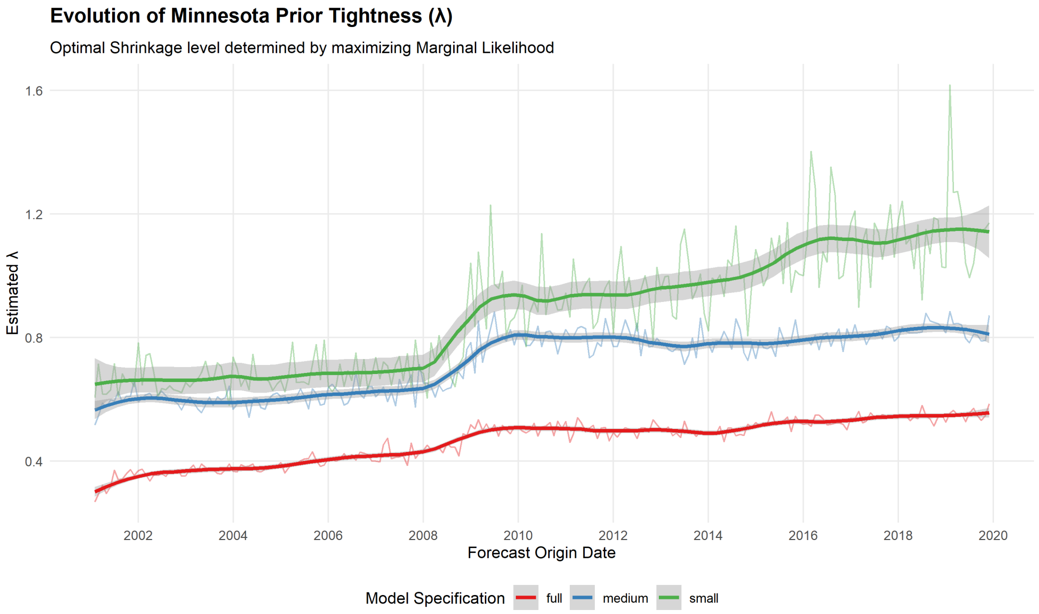


Figure 7: Evolution of hierarchical tightness parameter λ over forecast origins

Notes: The figure plots posterior means of λ at each recursive forecast origin from 2001M1 to 2019M11 for each model specification. λ is elevated during the 2008–2010 crisis window and increases further in the late sample. Source: `results/forecasts/hyperparameters_evolution.csv`.

4.6 Robustness

Two robustness exercises vary (i) the lag length from $p = 12$ to $p = 6$ and (ii) the initial training window endpoint from 2000M12 to 1995M12, keeping the rest of the design unchanged. The qualitative implications are stable. The medium specification remains the strongest performer for industrial production at short horizons, and the full specification remains competitive for inflation. The alternative initial-window design yields the lowest twelve-month inflation RMSFE for the full model (1.286). Appendix Table **??** and Appendix Figure **??** summarize these results.

5 Limitations and future research directions

5.1 Data and identification boundaries

Our analysis uses pseudo out-of-sample forecasts constructed from final-vintage macroeconomic data, not real-time vintages that forecasters would have actually observed. This choice simplifies the analysis and focuses attention on the information content of various data sources, but it sidesteps the practical challenge of nowcasting and data revision that practitioners face. A natural extension would be to re-implement this analysis using FRED-RTDF real-time data, which would reveal whether sentiment's predictive value survives the revision process—i.e., whether sentiment indices themselves are robust to later revision.

The paper is deliberately descriptive and does not attempt to identify causal relationships between sentiment and macro outcomes. Consumer sentiment and macroeconomic conditions are mutually endogenous: households' sentiment responds to current conditions (employment, inflation expectations, asset prices), and in turn, sentiment-driven changes in consumption and savings affect output and inflation. A structural VAR exercise (estimating causal impulse responses via sign or zero restrictions) is beyond the paper's scope, but it would be a valuable complement to clarify the direction of causality and the quantitative magnitude of sentiment's causal effect.

5.2 Model specification and functional form

The paper estimates a linear BVAR on all three information sets. Inflation and sentiment may be related through nonlinear channels: for instance, sentiment's predictive content might be stronger during crisis periods (high volatility, low sentiment) than during calm periods. A time-varying parameter VAR (TV-BVAR) or a model with regime-switching could capture this richer dynamic. Similarly, we do not explore whether sentiment is better measured by decomposing the Michigan index into sub-components (current vs. expected conditions) or by combining sentiment with alternative confidence measures (e.g., the Conference Board consumer confidence index).

The evaluation-scale transformation (cumulative growth over h periods from a fixed origin) is standard for forecast evaluation but may mask phenomena visible at other horizons. For instance, one-period-ahead growth-rate forecasts (as opposed to cumulative h-period changes) might reveal different roles for sentiment.

5.3 Limitations of the CG diagnostic for model-based forecasts

The Coibion-Gorodnichenko regression was originally developed to diagnose biases in survey expectations, where $\beta > 0$ can be interpreted as information rigidity or rational inattention by households. When applied to a VAR forecasting model, the interpretation is less direct: β measures the model's internal consistency in updating, not a structural behavioral phenomenon. A $\beta = 2.4$ coefficient at $h = 1$ for the small model means that the model tends to under-weight forecast revisions relative to the magnitude needed to eliminate subsequent errors, but this may reflect not an economic irrationality but a prior specification (the Minnesota prior may be too tight at short horizons) or a genuinely persistent signal that takes time to be incorporated.

5.4 Concrete proposals for extension

1. **Real-time data and nowcasting.** Repeat the analysis using FRED-RTDF real-time data vintages at forecast origin t, incorporating realistic delays and revisions. Assess whether sentiment's predictive value is diminished by data uncertainty.

2. **Time-varying and nonlinear structures.** Extend to TV-BVAR or Markov-switching BVAR to test whether sentiment's role varies across regimes (e.g., stronger during crisis periods or high-uncertainty environments).

3. **Sentiment decomposition.** Decompose the Michigan sentiment index into its major sub-components (current conditions vs. expectations) and evaluate their independent predictive contributions. Explore whether the expectations sub-component better predicts long-horizon inflation.

4. **Multivariate sentiment measures.** Combine the Michigan index with other sentiment indicators (Conference Board, stock market-based measures, news-based indices) and evaluate whether a factor model of sentiment improves predictions.

5. **Structural identification.** Estimate sign-restricted VAR IRFs to identify the causal response of inflation and production to a structural sentiment shock, holding constant the responses to other shocks.

6. **Comparative evaluation against production models.** Benchmark the BVAR's forecasts against professional forecasts from the Survey of Professional Forecasters (SPF) and other real-world prediction systems to assess practical competitive advantage.

6 Conclusion

This paper evaluates the incremental role of consumer sentiment in macro forecasting within a transparent, nested-information-set BVAR horse race, and complements accuracy comparisons with a forecast-revision diagnostic following ?. The evidence is consistent with a disciplined message: adding financial variables helps some short-horizon forecasts (especially for industrial production), but sentiment adds little incremental improvement in point-forecast RMSFE once financial variables are included; for inflation at long horizons, the baseline macro specification performs best in this sample.

The revision diagnostic shows that inflation forecasts exhibit short-horizon underreaction and long-horizon coefficients near zero or slightly negative, and that richer information sets move long-horizon coefficients toward the rational-expectations benchmark. Because the models are nested, I treat standard equal-accuracy tests as suggestive and report Clark–West MSPE-adjusted tests for nested comparisons **?**, emphasizing magnitudes and stability rather than sharp claims about statistical dominance **?**. The main limitation is interpretational: both RMSFE differences and CG coefficients are descriptive summaries of forecast performance rather than causal effects of sentiment. Future work using real-time vintages, broader sentiment measures, and structural identification would sharpen the economic interpretation.

A Additional figures and robustness

Nested-model forecast accuracy: Clark–West tests. Table **??** reports Clark–West MSPE-adjusted tests for nested model comparisons (Small vs. Medium; Medium vs. Full) at horizons $h \in \{1, 3, 12\}$; one-sided p-values correspond to the alternative that the larger model improves MSPE. This robustness addresses the nonstandard behavior of standard equal-accuracy tests under nesting.

Table 4: Clark–West (2007) MSPE-Adjusted Tests for Nested Models

Smaller	Larger	variable	horizon	t-stat	p-value	N	NW lag
Small	Medium	CPI	h=1	3.312***	0.001	227.000	1.000
Small	Medium	CPI	h=3	2.405***	0.008	225.000	3.000
Small	Medium	CPI	h=12	-0.063	0.525	216.000	12.000
Small	Medium	INDPRO	h=1	3.211***	0.001	227.000	1.000
Small	Medium	INDPRO	h=3	2.387***	0.009	225.000	3.000
Small	Medium	INDPRO	h=12	2.452***	0.008	216.000	12.000
Medium	Full	CPI	h=1	-1.146	0.874	227.000	1.000
Medium	Full	CPI	h=3	-0.325	0.627	225.000	3.000
Medium	Full	CPI	h=12	0.742	0.230	216.000	12.000
Medium	Full	INDPRO	h=1	0.107	0.458	227.000	1.000
Medium	Full	INDPRO	h=3	0.057	0.477	225.000	3.000
Medium	Full	INDPRO	h=12	0.253	0.400	216.000	12.000

Notes: Clark–West (2007) MSPE-adjusted test for equal forecast accuracy in nested models.
For smaller-model forecast error $e_{1t} = y_t - f_{1t}$ and larger-model error $e_{2t} = y_t - f_{2t}$, the adjusted loss differential is
$d_t = e_{1t}^2 - \left(e_{2t}^2 - (f_{2t} - f_{1t})^2\right)$. The test regresses d_t on a constant.
Newey–West HAC standard errors use lag truncation equal to the forecast horizon (overlap adjustment).
One-sided p-values reported for the alternative that the larger model improves MSPE.
*** p<0.01, ** p<0.05, * p<0.1

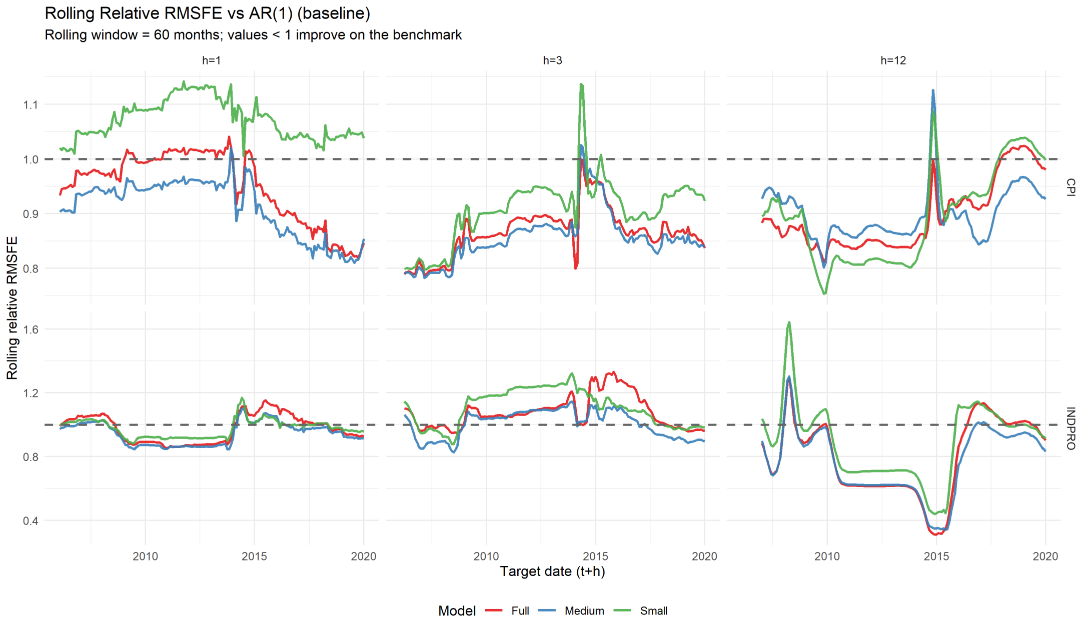


Figure 8: Rolling relative RMSFE versus AR(1) benchmark

Notes: Rolling-window relative RMSFEs (window length 60 months). Values below one indicate improvement over the recursively estimated AR(1) benchmark.

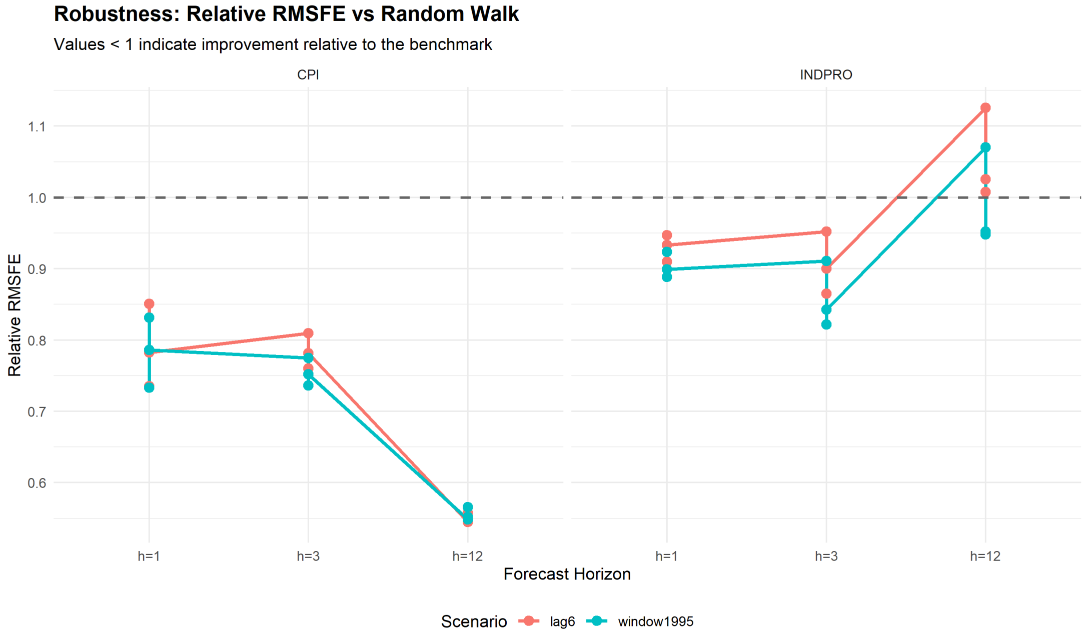


Figure 9: Robustness: relative RMSFE versus no-change benchmark

Notes: Relative RMSFEs under alternative lag length ($p = 6$) and an earlier initial training window end date (1995M12). Values below one indicate improvement over the no-change benchmark.

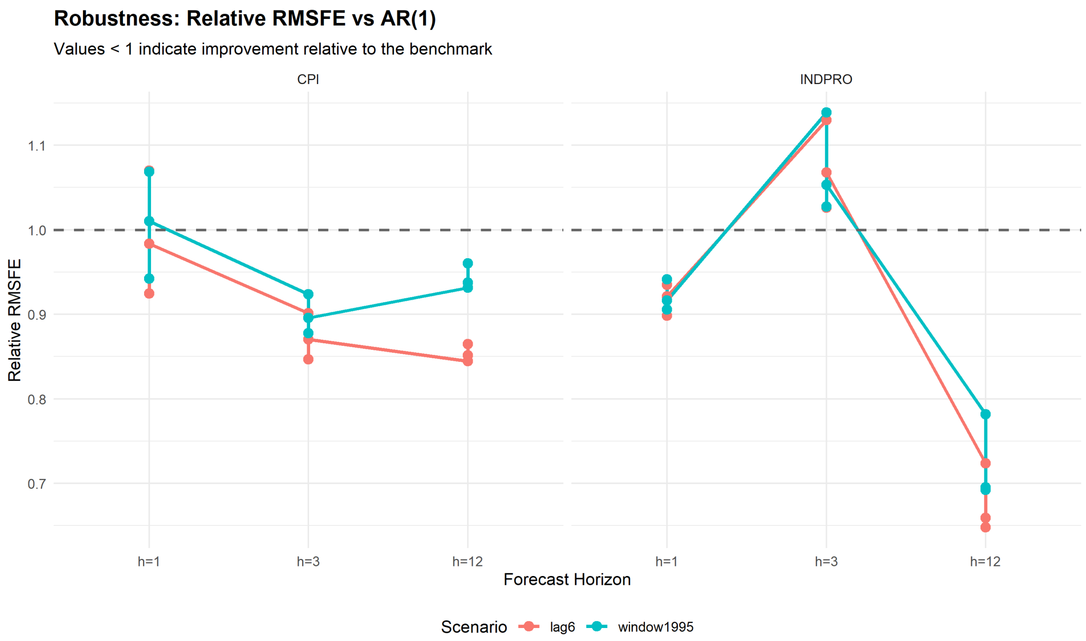


Figure 10: Robustness: relative RMSFE versus AR(1) benchmark

Notes: Relative RMSFEs under robustness scenarios, reported against the recursively estimated AR(1) benchmark.

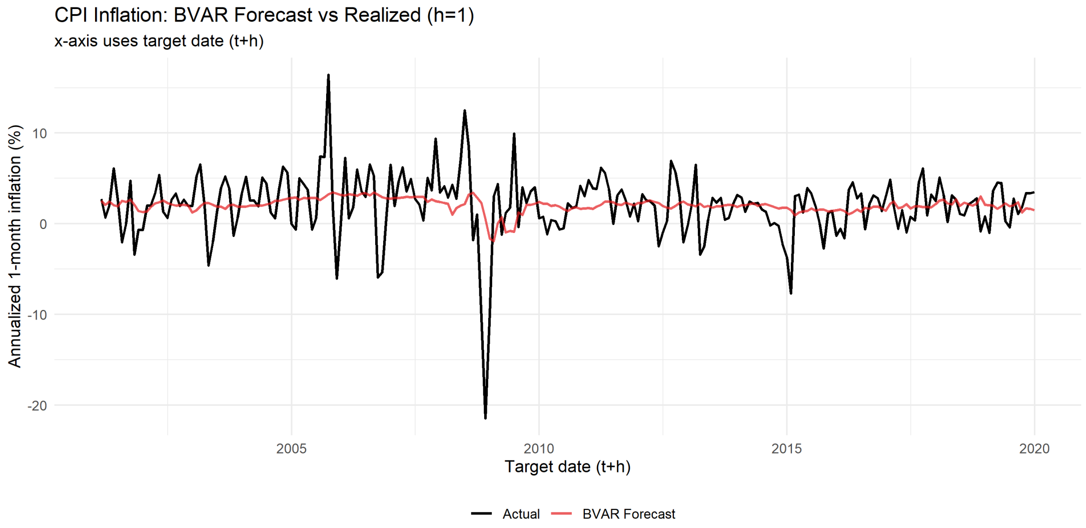


Figure 11: CPI inflation: BVAR forecast versus realized ($h = 1$)

Notes: The x-axis uses the target date ($t + h$). The plotted forecast is the model-implied predictive mean from the baseline specification.

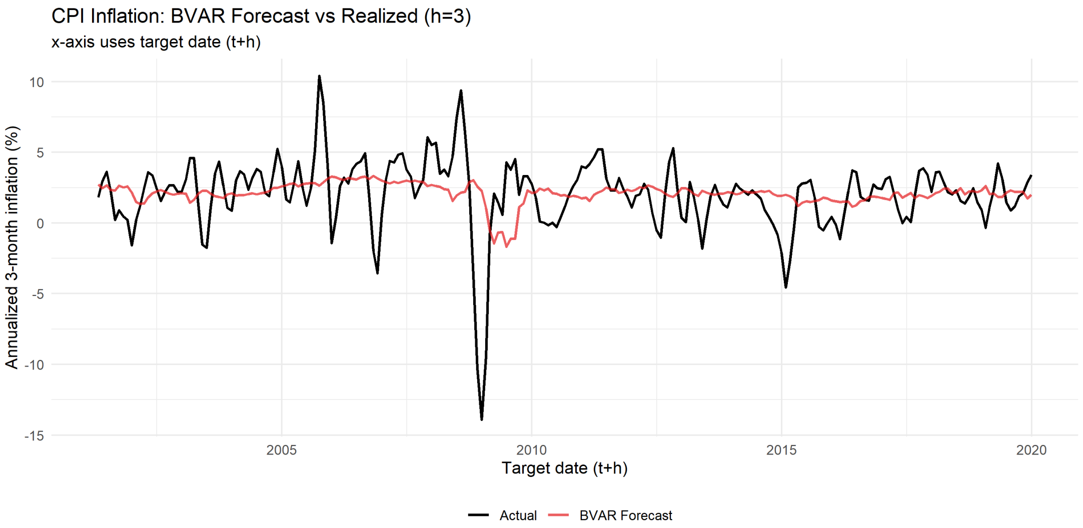


Figure 12: CPI inflation: BVAR forecast versus realized ($h = 3$)

Notes: The x-axis uses the target date ($t + h$). The plotted forecast is the model-implied predictive mean from the baseline specification.

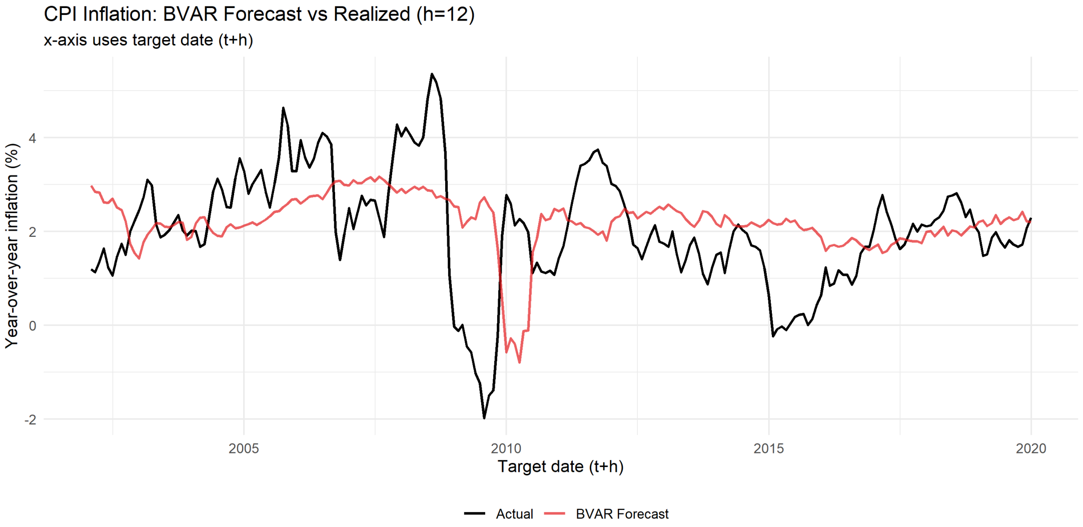


Figure 13: CPI inflation: BVAR forecast versus realized ($h = 12$)

Notes: The x-axis uses the target date ($t + h$). The plotted forecast is the model-implied predictive mean from the baseline specification.

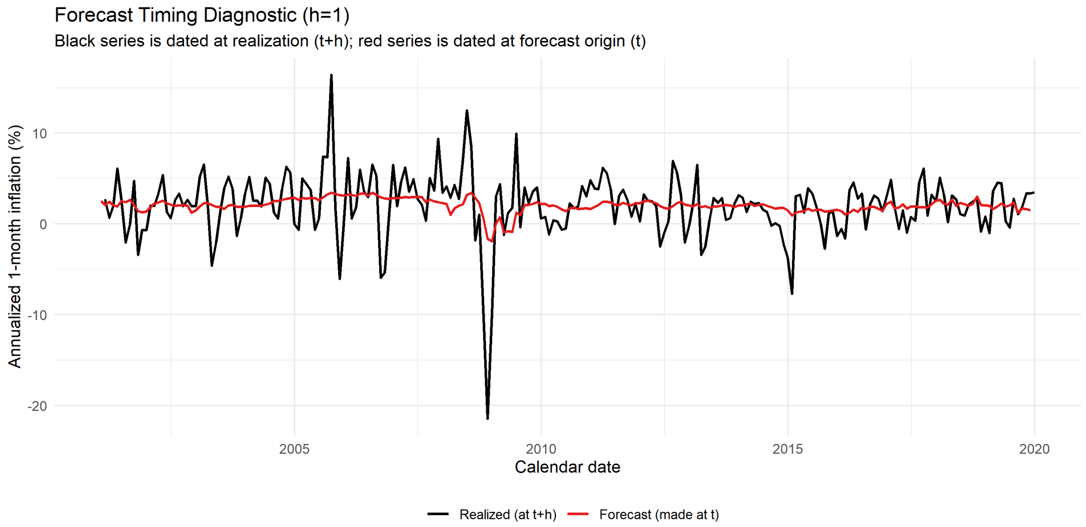


Figure 14: Forecast timing diagnostic ($h = 1$)

Notes: The black series is dated at the realization ($t + h$); the red forecast series is dated at the forecast origin (t).

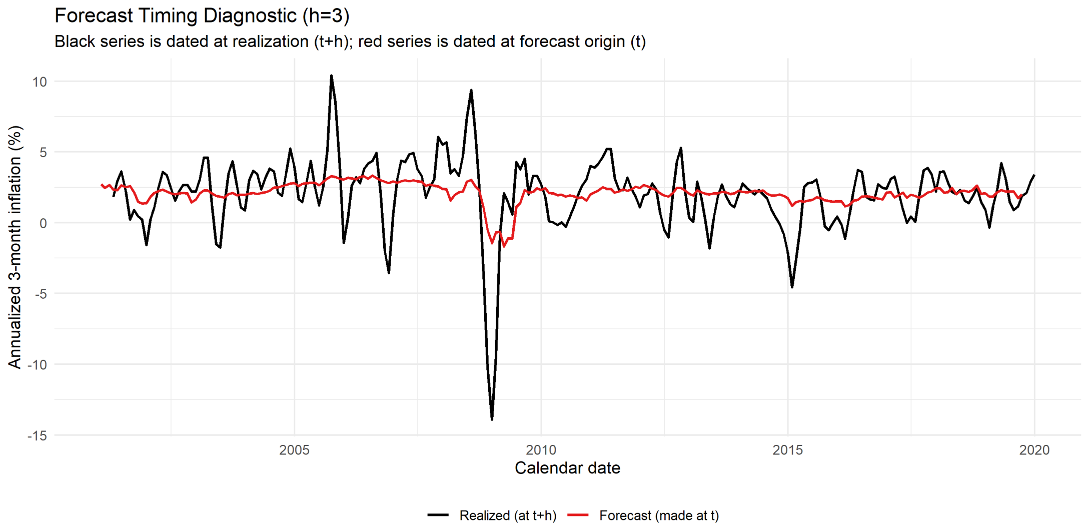


Figure 15: Forecast timing diagnostic ($h = 3$)

Notes: The black series is dated at the realization ($t + h$); the red forecast series is dated at the forecast origin (t).

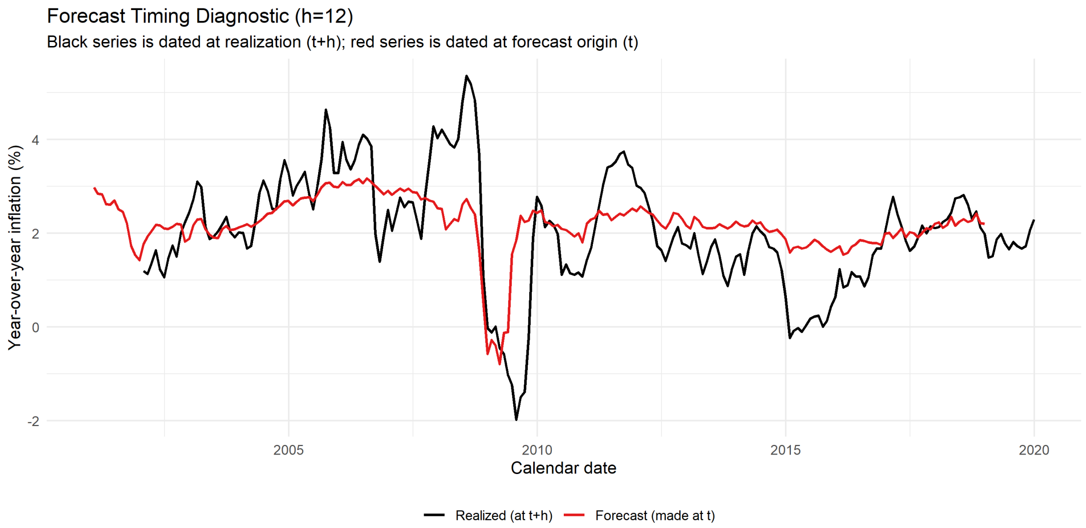


Figure 16: Forecast timing diagnostic ($h = 12$)

Notes: The black series is dated at the realization $(t + h)$; the red forecast series is dated at the forecast origin (t).

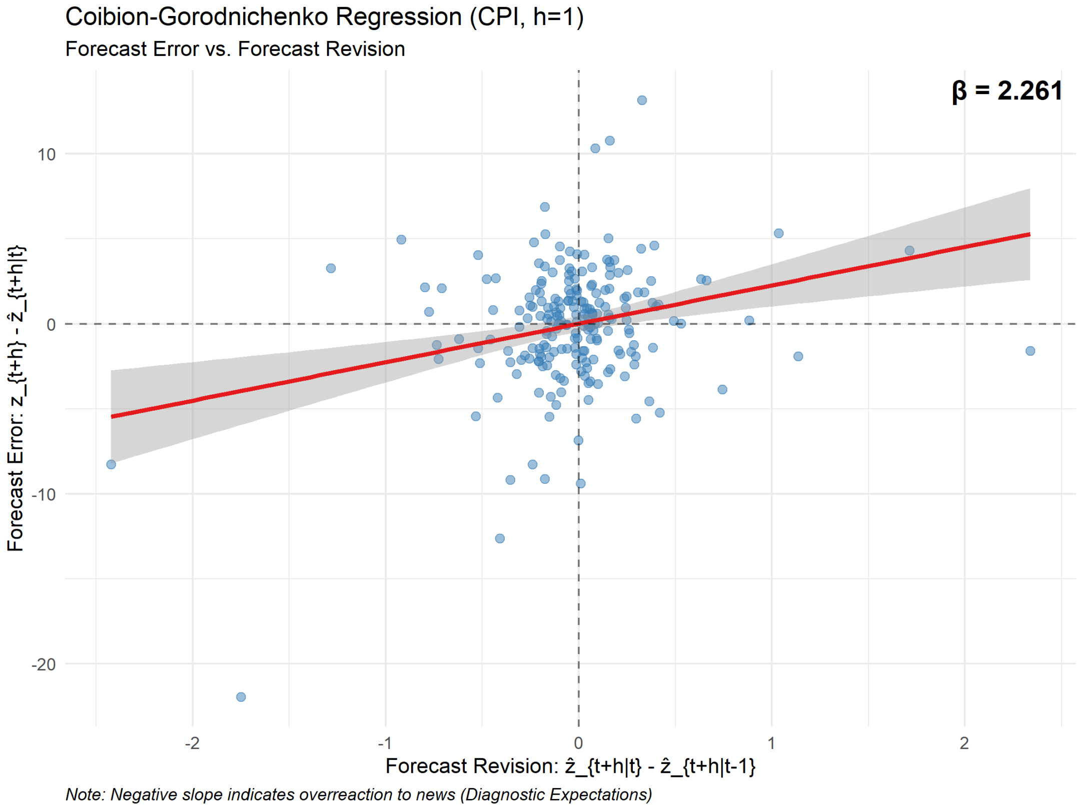
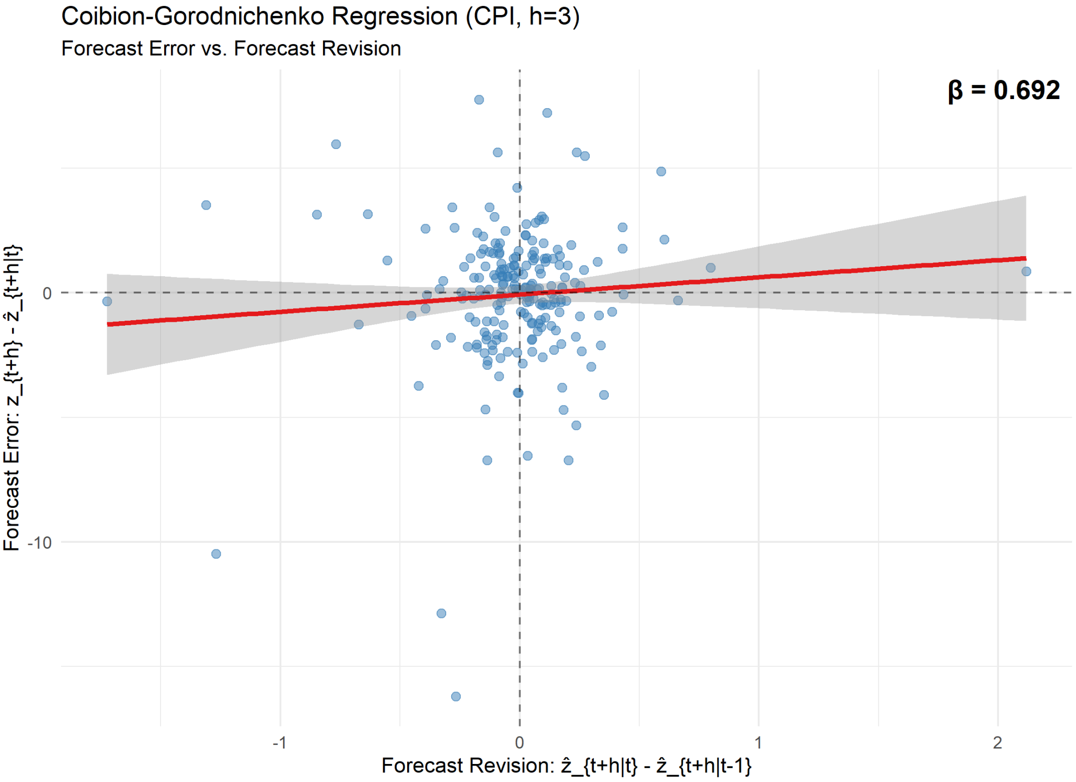


Figure 17: Revision diagnostic scatter: CPI ($h = 1$)

Notes: Scatter of forecast errors against forecast revisions for CPI inflation in the baseline design. The fitted line corresponds to the **?** regression.



Note: Negative slope indicates overreaction to news (Diagnostic Expectations)

Figure 18: Revision diagnostic scatter: CPI ($h = 3$)

Notes: Scatter of forecast errors against forecast revisions for CPI inflation in the baseline design. The fitted line corresponds to the **?** regression.

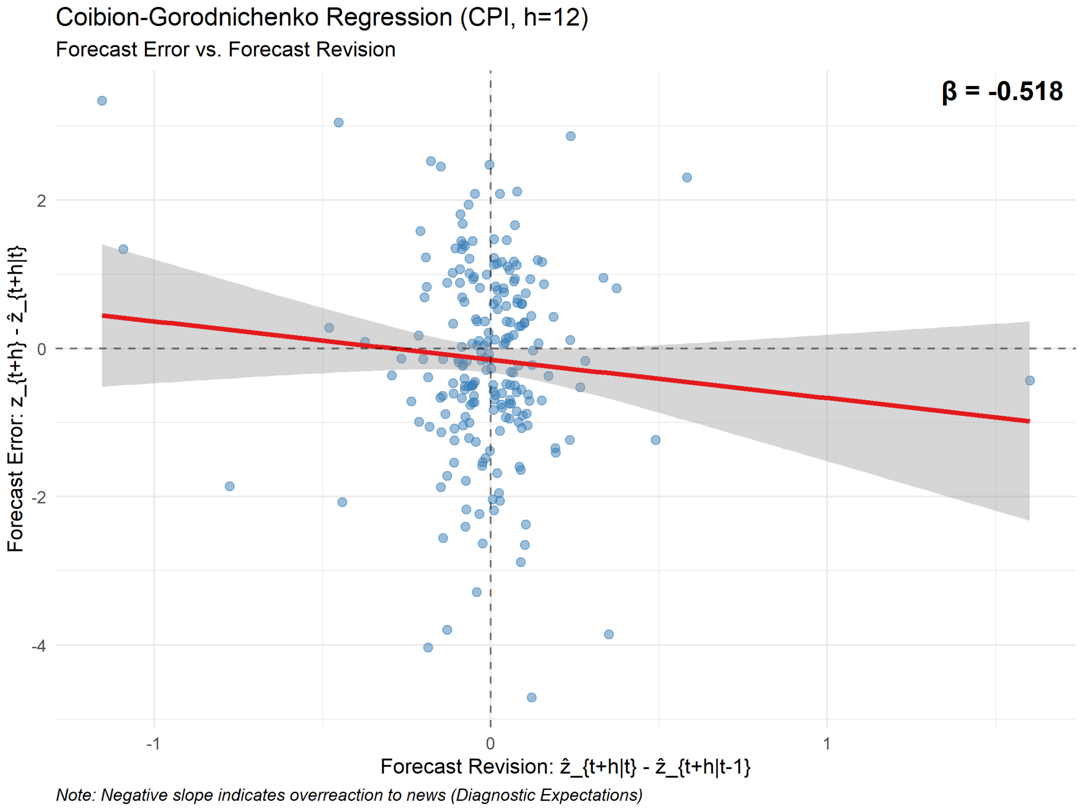


Figure 19: Revision diagnostic scatter: CPI ($h = 12$)

Notes: Scatter of forecast errors against forecast revisions for CPI inflation in the baseline design. The fitted line corresponds to the **?** regression.

Table 5: Robustness: RMSFEs under alternative lag length and training window

Scenario	Model	Target	$h = 1$	$h = 3$	$h = 12$
$p = 6$	Small	CPI	3.451	2.639	1.291
$p = 6$	Medium	CPI	3.445	2.641	1.294
$p = 6$	Full	CPI	3.529	2.673	1.342
$p = 6$	Small	INDPRO	7.582	5.410	4.825
$p = 6$	Medium	INDPRO	7.336	5.033	4.621
$p = 6$	Full	INDPRO	7.494	5.173	4.572
Initial window ends 1995M12	Small	CPI	3.231	2.448	1.323
Initial window ends 1995M12	Medium	CPI	3.216	2.446	1.325
Initial window ends 1995M12	Full	CPI	3.267	2.431	1.286
Initial window ends 1995M12	Small	INDPRO	7.329	5.204	4.802
Initial window ends 1995M12	Medium	INDPRO	7.077	4.803	4.590
Initial window ends 1995M12	Full	INDPRO	7.218	4.930	4.453

Notes: Values are taken from `results/robustness/lag6/tables/rmsfe_results.csv` and `results/robustness/window1995/tables/rmsfe_results.csv`.