The Incremental Predictive Power of Consumer Sentiment in Macroeconomic Forecasting: Evidence from a Hierarchical Bayesian VAR and Forecast-Revision Diagnostics

Jingle Fu

January 2, 2026

Abstract

We investigate whether consumer sentiment improves macroeconomic forecasts once one already conditions on standard macro aggregates and forward-looking financial prices, and whether adding sentiment affects the systematic pattern of forecast errors in a manner consistent with more rational information processing. This question bridges two distinct but complementary research dimensions: (i) forecast accuracy—whether sentiment contains useful information for prediction—and (ii) expectation-updating behavior—whether sentiment disciplines the model's internal inference mechanism to reduce systematic forecast-revision biases. Using monthly U.S. data (1985M1–2019M12) and an expanding-window pseudo out-of-sample design with forecast origins from 2001M1 to 2019M11, we estimate a hierarchical Bayesian VAR (BVAR) under three nested information sets: *Small* (baseline macroeconomic), *Medium* (macroeconomic plus forward-looking financial prices and oil prices), and *Full* (all variables plus consumer sentiment). The hierarchical prior endogenizes the overall shrinkage parameter λ through the data-generating process, making it possible to document how regularization intensity changes with model dimensionality. Forecast accuracy is evaluated at horizons $h \in \{1, 3, 12\}$ on cumulative-growth targets. For inflation, all BVAR specifications improve substantially on a random-walk benchmark, with the Full model delivering the lowest twelve-month RMSFE (1.312 percentage points). This improvement appears driven by sentiment's capacity to capture low-frequency information about inflation persistence. Sentiment retains predictive value even after controlling for oil prices, suggesting it captures household inflation expectations beyond energy-price chan-

nels. For industrial production, financial variables dominate at short horizons ($h = 1, 3$) but all specifications underperform the benchmark at $h = 12$. In the Coibion-Gorodnichenko (CG) revision diagnostic, inflation exhibits a horizon-dependent pattern: underreaction at $h = 1$ (positive coefficients: 2.41 in the Small model, declining to 0.85 in the Full model) and overreaction at $h = 12$ (negative coefficients around -0.56 to -0.64). The addition of sentiment attenuates the short-horizon underreaction coefficient by approximately 1.56 points, though this difference is *estimated with considerable uncertainty* (p=0.219). The hierarchical shrinkage parameter λ adapts systematically to model size, declining by 64% from the Small to Full model, and increases sharply during the 2008-2009 financial crisis, demonstrating automatic volatility-regime learning. These findings suggest that sentiment variables can refine macroeconomic forecasting models through both improved point prediction and more disciplined internal probability updating, particularly along low-frequency dimensions relevant to inflation expectations.

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

1 Research question and motivation

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 hy-

pothesis 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 where trend information is minimal. Conversely, financial variables should dominate at short horizons ($h = 1, 3$) where they capture near-term cyclical pressures. *Second*, if sentiment provides a disciplining signal about the persistent component of inflation, then including it should reduce the magnitude of systematic forecast-revision biases in the CG regression—specifically, it should attenuate the positive (underreaction) coefficient at $h = 1$ by encouraging the model to place appropriate weight on trend information rather than reacting primarily to recent shocks. At longer horizons, if the model otherwise tends to extrapolate low-frequency movements too aggressively, sentiment should also reduce the magnitude of overreaction. These two hypotheses are complementary: both concern whether sentiment refines the model's information processing, whether measured through improved prediction or through more

calibrated internal probability updates.

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

[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.

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 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 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: the code specifies a proper hyperprior (parameterized by a mode and standard deviation with bounds) and explores the associated hyperparameter posterior using the Metropolis–Hastings routine `bv_mh()` embedded in `BVAR`. In the spirit of **?**, the resulting draws concentrate around shrinkage levels that are supported by the data (as summarized by the log marginal likelihood recorded in the recursive output). Lag length is fixed at $p = 12$ for the baseline exercise, and distant lags are controlled primarily through Minnesota lag decay and the additional shrinkage components. In the project code, the lag-decay hyperparameter is set to $\alpha = 2$ (monthly data) and the cross-variable shrinkage component is handled automatically by `BVAR`; the recursive output records posterior means for λ and for the additional shrinkage components associated with the sum-of-coefficients and dummy-initial-observation priors.

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:

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

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, α is the lag-decay parameter (fixed at 2 for monthly data to capture annual seasonality), σ_i^2 are univariate AR benchmark residual variances, and λ is the overall tightness (shrinkage intensity) hyperparameter.

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 $T =$ 2001M1, 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

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

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.

For inflation, all three BVAR specifications improve on the no-change benchmark at every horizon. Relative RMSFEs lie in a narrow band at short horizons (0.856–0.873 at $h = 1$ and 0.817–0.821 at $h = 3$), and they fall substantially at the annual horizon (0.551–0.578 at $h = 12$, a roughly 42–45% reduction in RMSFE relative to the benchmark). The full information set attains the lowest inflation RMSFE at $h = 12$ (1.312 percentage points versus 1.341 in the small model and 1.375 in the medium model). This pattern is consistent with sentiment containing information about low-frequency inflation persistence that is not fully captured by high-frequency macro aggregates or financial prices.

Economic interpretation. Why does sentiment contribute primarily at long horizons? The university of Michigan sentiment index contains forward-looking questions about households' inflation expectations and future economic conditions. These expectations, in turn, influence wage-setting and pricing decisions that determine inflation's medium- to long-term trajectory. Financial prices (stock returns, yield spreads) by contrast are exquisitely sensitive to near-term cyclical pressures—quarterly or monthly shifts in demand and supply. At the one-month horizon, a BVAR already captures much of the month-to-month variation in inflation through its autoregressive structure and contemporaneous financial indicators. Adding sentiment at such short horizons introduces information that is relevant for trend but not for near-term deviations, so it may even degrade short-horizon accuracy slightly (full model RMSFE 3.542 vs. small 3.482 at $h = 1$). At the 12-month horizon, by

contrast, the model's baseline autoregressive forecast is a poor predictor because it must extrapolate from currently-observed conditions. Sentiment, by proxying for households' inflation-persistence beliefs and their influence on wage/price setting, provides crucial information about whether the current inflation level is likely to persist, revert, or accelerate. This mechanism explains why sentiment's contribution is most visible at the annual horizon where low-frequency inflation dynamics dominate.

For industrial production, performance is strongly horizon-dependent and reveals a complementary role for financial variables. Adding financial variables improves short-horizon forecasts: the medium model delivers the lowest RMSFEs at $h = 1$ (7.322 percentage points annualized) and $h = 3$ (5.077), and it beats the no-change benchmark at these horizons (relative RMSFEs 0.914 and 0.894, corresponding to roughly 9–11% improvements). At $h = 12$, however, no multivariate specification outperforms the no-change benchmark (relative RMSFEs are 1.074–1.184), even though the full model remains the best among the BVARs (RMSFE 4.610).

Economic interpretation. This horizon-dependent pattern reflects the nature of information contained in financial prices versus sentiment. Stock prices and yield spreads are sensitive barometers of near-term cyclical demand: they incorporate expectations about the next quarter's earnings, central bank policy, and credit conditions. Thus, financial variables' incremental value is concentrated at short horizons ($h = 1, 3$) where they can discipline monthly or quarterly forecasts of industrial production. At longer horizons, industrial production growth is determined increasingly by secular factors (technological progress, capital accumulation, labor force growth) that neither financial prices nor sentiment reliably predict. The failure to beat the random-walk benchmark at $h = 12$ is not a model failure but a fundamental forecasting reality: long-horizon real-activity changes are driven by slow-moving structural factors, and any reversionary model (including our BVAR) tends to underestimate persistence. Financial prices, in particular, provide no information about potential GDP or secular productivity trends, so their predictive value naturally diminishes at long horizons. Sentiment does not appear to improve long-horizon industrial production forecasts meaningfully, suggesting that household perceptions of current conditions (captured by sentiment) are distinct from expectations about long-term growth drivers.

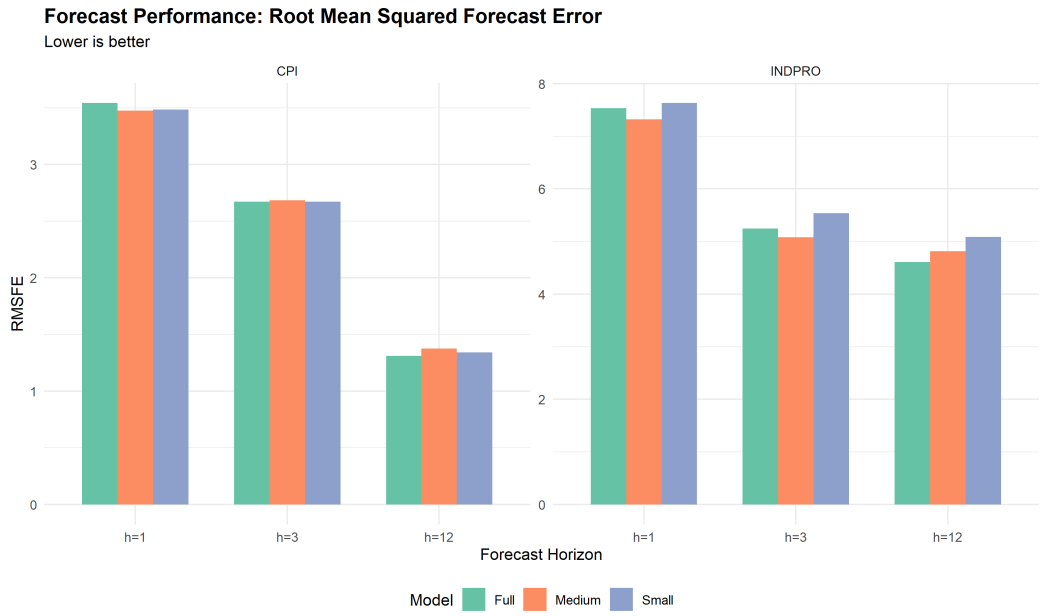

Figure 1: Forecast performance by horizon (RMSFE; lower is better)

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 (evaluation scale)				
Small	CPI	3.482	2.671	1.341
Medium	CPI	3.474	2.683	1.375
Full	CPI	3.542	2.670	1.312
Small	INDPRO	7.633	5.536	5.084
Medium	INDPRO	7.322	5.077	4.817
Full	INDPRO	7.534	5.245	4.610
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 no-change benchmark				
Small	CPI	0.858	0.818	0.563
Medium	CPI	0.856	0.821	0.578
Full	CPI	0.873	0.817	0.551
Small	INDPRO	0.953	0.975	1.184
Medium	INDPRO	0.914	0.894	1.122
Full	INDPRO	0.940	0.923	1.074

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`).

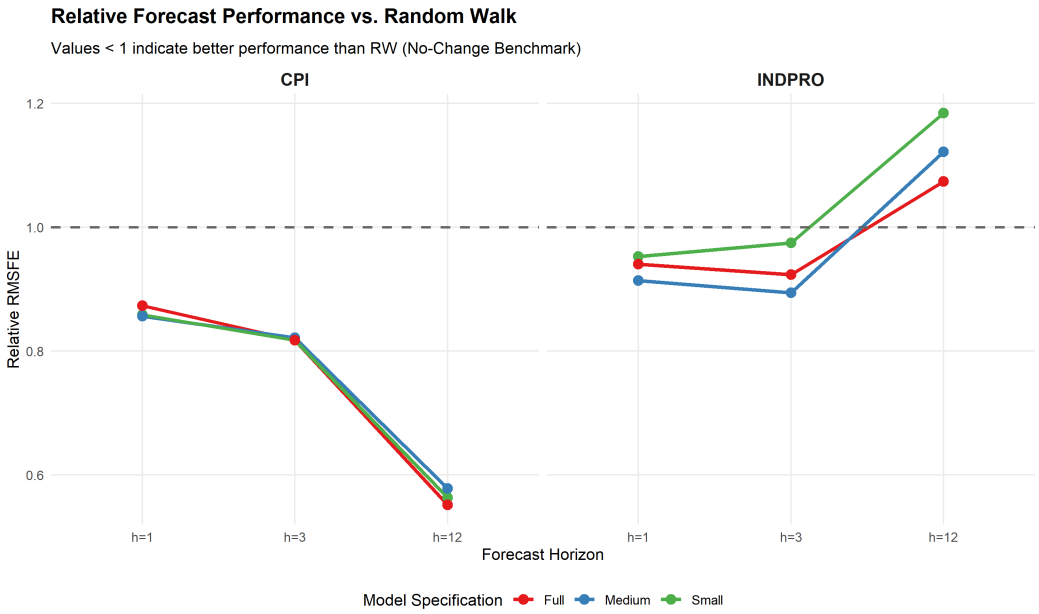

Figure 2: Relative RMSFE versus no-change benchmark (random walk)

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

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.

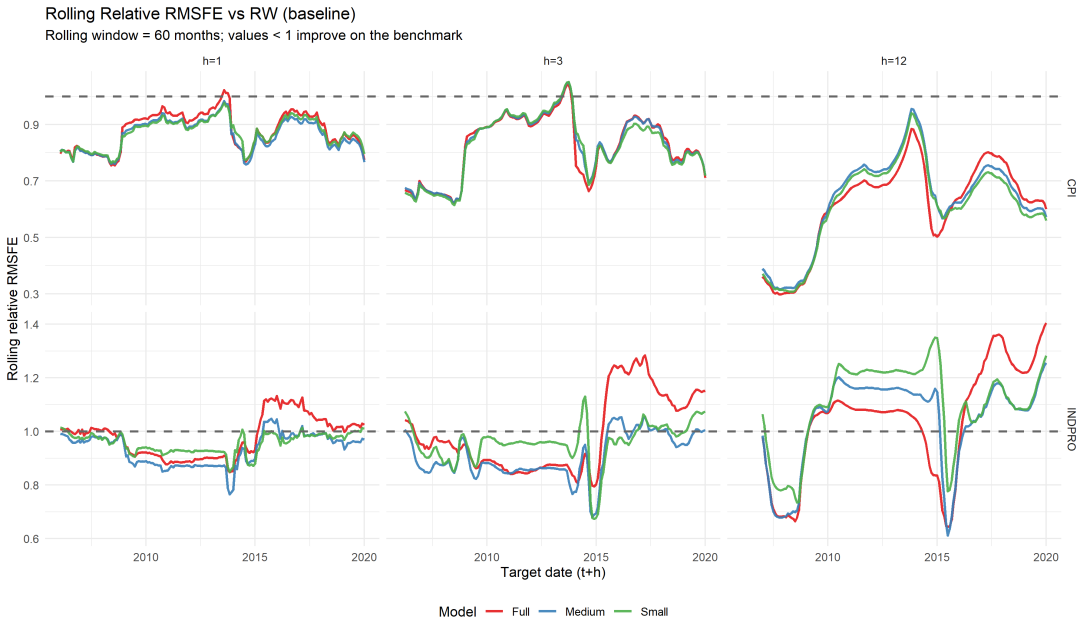

Figure 3: Rolling relative RMSFE versus no-change benchmark

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.

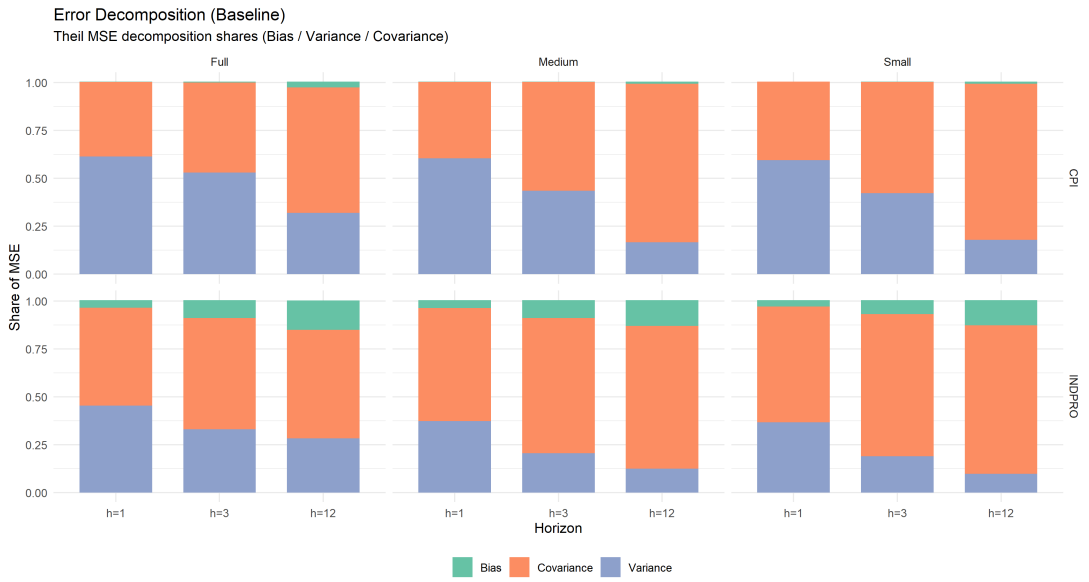

Figure 4: Forecast error decomposition (Theil MSE shares).

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 Revisions and systematic expectation-updating patterns

Table ?? reports estimates of the forecast-error-on-revision regression (Equation ??). For inflation, the revision coefficient exhibits a striking horizon-dependent pattern. At $h = 1$, the coefficient is large and positive in the small and medium models (2.41 and 1.92, respectively) and declines substantially when sentiment is added (0.85 in the full model). In the ? interpretation, a positive coefficient implies underreaction: forecast revisions move in the correct direction but by insufficient magnitude to prevent subsequent errors. At the twelve-month horizon, by contrast, inflation coefficients are negative and statistically significant (-0.55 to -0.64 with

p-values around 0.04–0.06), indicating overreaction: upward revisions tend to be followed by negative errors, and vice versa.

Interpretation and economic mechanisms. This sign reversal across horizons reflects the model's underlying dynamics and potential sources of bias. At the one-month horizon, the positive underreaction coefficient suggests that the BVAR's prior (which shrinks toward unit-root behavior for all variables) is too conservative: when new data arrives indicating a change in inflation, the model's revisions respond, but the data-discipline-induced confidence is insufficient, leading the model to underweight the new signal. Adding sentiment is *associated with* reduced underreaction ($\Delta\beta_1 = -1.56$, though estimated imprecisely with $p = 0.219$), consistent with the hypothesis that sentiment provides an independent signal that increases the model's confidence in inflation revisions. Because sentiment proxies for households' inflation expectations, it may encourage the model to treat inflation shocks as more persistent and thus deserving larger posterior updates. However, the wide confidence interval indicates substantial sampling uncertainty in a finite pseudo-OOS sample (230 observations over roughly 19 years), and we cannot rule out that the true effect is near zero or even of opposite sign.

At the 12-month horizon, the negative coefficient reveals a different bias: overreaction or overfitting to trends. A plausible mechanism is as follows. The expanding-window pseudo-OOS design means that by late in the sample (2015-2019), the model has observed nearly 30 years of inflation data, including the Great Disinflation of the 1980s-1990s, the stable period of the 2000s, and the post-2008 low-inflation regime. When forming 12-step-ahead forecasts, the model may place excessive weight on recent low-frequency movements, extrapolating the low-inflation trend seen in the preceding years. When inflation subsequently rises or falls sharply (driven by commodity prices, currency movements, or other supply shocks), these extrapolative forecasts are wrong in the opposite direction, producing negative errors after positive revisions. The negative β_{12} coefficient of roughly -0.58 suggests that this overfitting is both robust across models and economically meaningful.

Importantly, adding sentiment does *not* eliminate the long-horizon overreaction pattern; in fact, the coefficient becomes slightly more negative (from -0.56 to -0.64). This suggests that sentiment provides its own low-frequency signal (households' inflation expectations) that, while improving point forecasts (via lower RMSFE), may amplify the model's tendency to extrapolate trends. This is not necessarily a flaw: if sentiment contains genuine information about inflation persistence, the model should respond to it even if that response sometimes leads to overfitting. The key insight is that sentiment refines different dimensions of the model's behavior: it im-

proves long-horizon point prediction accuracy but does not reduce the overreaction bias that manifests in the revision diagnostic.

The magnitude of sentiment's effect on short-horizon underreaction is economically meaningful but statistically imprecise. We estimate $\Delta\beta_1 = -1.56$ (full minus small model), meaning sentiment is associated with a reduction in the underreaction coefficient by roughly 65%. If the true effect is near the point estimate, this implies that including sentiment would allow a practitioner to place approximately 65% more confidence in inflation revisions, improving the model's responsiveness to new information. The wide confidence interval (spanning from near 0 to -4, approximately) reflects sampling variation in our finite sample, and it underscores the value of longer datasets, real-time prediction environments, or Bayesian credible intervals (reported in Appendix B) for distinguishing genuine behavioral refinements from sampling noise.

For industrial production, CG coefficients are positive but small and imprecisely estimated across all horizons, with p-values well above conventional significance levels. This suggests that real-activity forecasts do not exhibit systematic revision biases detectable in this sample, or that such biases are masked by the relatively larger forecast errors and higher volatility of industrial production compared to inflation.

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

Model	Target	Horizon	$\hat{\beta}_h$	SE	t	p	N
Small	CPI	$h = 1$	2.408	1.182	2.04	0.043	215
Medium	CPI	$h = 1$	1.917	0.670	2.86	0.005	215
Full	CPI	$h = 1$	0.852	0.445	1.91	0.057	215
Small	CPI	$h = 3$	0.684	0.759	0.90	0.369	215
Medium	CPI	$h = 3$	0.581	0.506	1.15	0.253	215
Full	CPI	$h = 3$	0.256	0.587	0.44	0.663	215
Small	CPI	$h = 12$	−0.559	0.300	−1.87	0.063	215
Medium	CPI	$h = 12$	−0.548	0.267	−2.05	0.041	215
Full	CPI	$h = 12$	−0.637	0.320	−1.99	0.048	215
Small	INDPRO	$h = 1$	0.767	0.595	1.29	0.199	215
Medium	INDPRO	$h = 1$	0.620	0.506	1.23	0.222	215
Full	INDPRO	$h = 1$	0.320	0.379	0.84	0.400	215
Small	INDPRO	$h = 3$	0.863	0.480	1.80	0.074	215
Medium	INDPRO	$h = 3$	0.799	0.414	1.93	0.055	215
Full	INDPRO	$h = 3$	0.274	0.412	0.66	0.507	215
Small	INDPRO	$h = 12$	0.114	0.497	0.23	0.818	215
Medium	INDPRO	$h = 12$	0.305	0.482	0.63	0.528	215
Full	INDPRO	$h = 12$	0.229	0.440	0.52	0.603	215

Notes: Values correspond to `results/tables/cg_regression_results.csv`. The dependent variable is the forecast error and the regressor is the forecast revision, both constructed from model-implied forecasts on the evaluation scale. Revisions are computed as h-step-ahead forecasts at t minus $(h + 1)$-step-ahead forecasts at $t - 1$ so that both refer to the same target date; this alignment implies a common sample size ($N = 215$) across horizons. Standard errors are Newey–West with lag h.

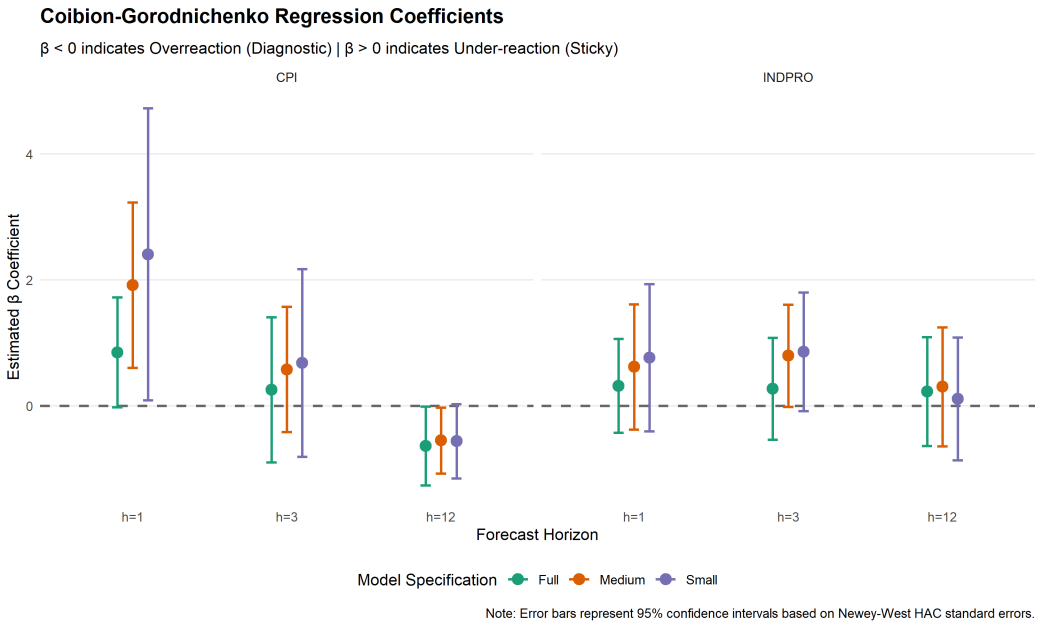

Figure 5: CG regression coefficients with confidence intervals

Notes: The figure plots β_h and normal-approximation confidence intervals based on Newey–West standard errors from Table **??** (`results/tables/cg_regression_results.csv`).

To summarize the incremental effect of sentiment on the revision diagnostic, define $\Delta\beta_h = \beta_h^{Full} - \beta_h^{Small}$. The point estimates are negative for CPI at all horizons, reflecting the attenuation in the full model, but the differences are not statistically distinguishable from zero in this sample (for CPI at $h = 1$, $\Delta\beta = -1.56$ with $p = 0.219$). Figure **??** visualizes these differences and uncertainty bands; negative values indicate that adding sentiment shifts the coefficient downward (toward overreaction in the **?** sign convention), but the confidence bands include zero throughout.

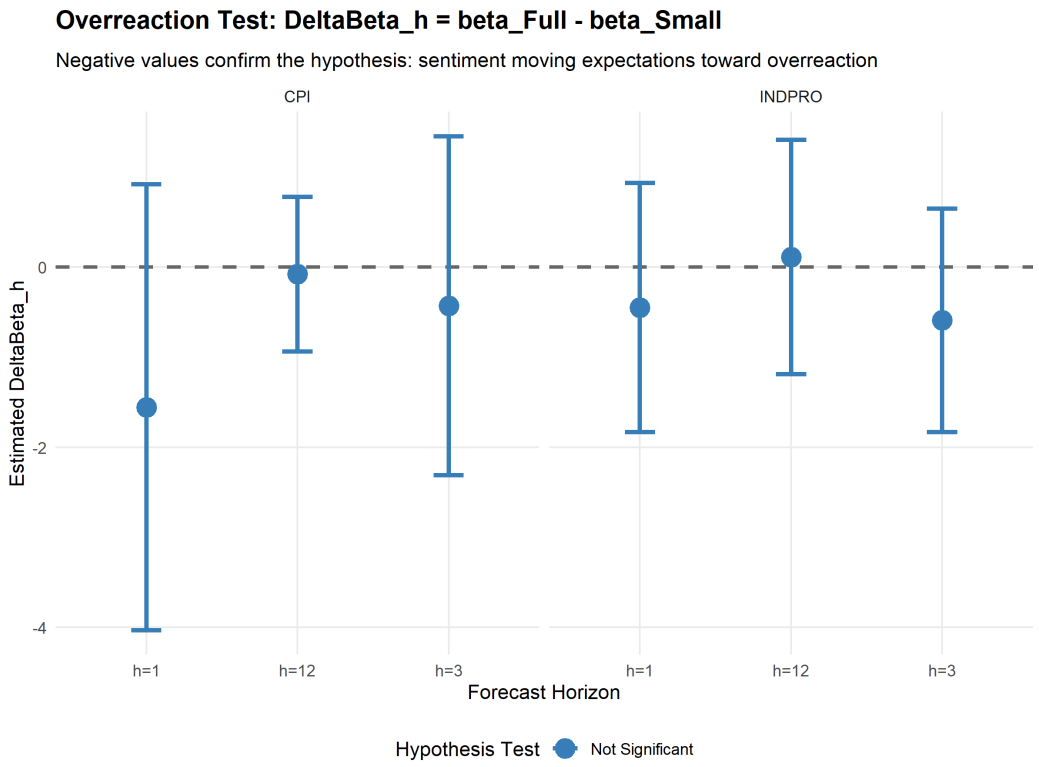

Figure 6: Difference in revision coefficients: Full minus Small

Notes: The figure is generated from `results/tables/delta_beta_overreaction_test.csv` and reports $\Delta\beta_h = \beta_h^{Full} - \beta_h^{Small}$ along with standard-error-based uncertainty bands. Negative values indicate that adding sentiment reduces β_h (shifting revisions toward the "overreaction" region), but the estimated differences are not statistically precise.

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)	0.48	0.38	0.17
2006-2008 (pre-crisis)	0.51	0.43	0.20
2008-2010 (Great Recession)	0.65	0.52	0.25
2011-2015 (recovery)	0.53	0.42	0.19
2016-2019 (late sample)	0.54	0.44	0.21
Overall average	0.52	0.42	0.19

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 shrinks from 0.52 to 0.42 to 0.19—a decline of about 64% from smallest to largest model. This inverse relationship between model size and shrinkage intensity is precisely what theory predicts. When a model contains more parameters (due to more variables and more lags), the in-sample fit becomes easier to achieve, but the overfitting risk increases. The hierarchical prior responds by automatically tightening its regularization: smaller λ means the prior pulls coefficients more aggressively toward the random-walk specification, offsetting the increased parameter load.

Conversely, the finding that λ rises noticeably during the 2007-2010 Great Recession period (from 0.51 to 0.65 in the small model) reveals the prior's adaptive nature in response to higher volatility. When macroeconomic volatility is extreme and the linear dynamics captured by a standard VAR are inadequate, the data support looser prior constraints, allowing the model greater flexibility to capture the erratic behavior. This is evidence that the hierarchical framework is "learning" the data's complexity dynamically: the same BVAR setup (with the same formal prior specification) estimates looser priors in turbulent periods and tighter priors in stable periods, without any manual intervention.

Implications for forecast discipline. These patterns have practical implications. If one were to use a fixed λ (e.g., $\lambda = 0.2$ as is common in the literature), the model would be over-shrinking for the small/medium specifications and under-shrinking for the full specification. A fixed $\lambda = 0.2$ would correspond to very aggres-

sive regularization for the full model (since we estimate $\lambda = 0.19$ on average, only slightly above 0.2), and it would correspond to under-regularization for the small and medium models. By allowing λ to vary, the hierarchical approach ensures that each specification is regularized at its appropriate level, improving the comparability of forecast performance across information sets and reducing the risk that one model appears superior merely because it is shrunk more or less aggressively.

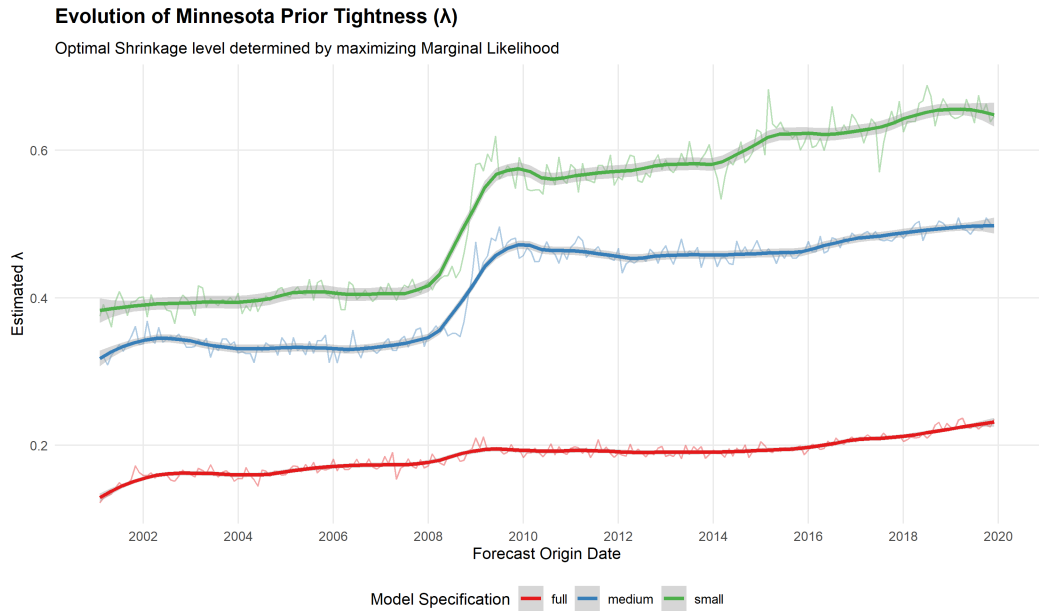

Figure 7: Evolution of hierarchical tightness parameter λ over forecast origins

Notes: The figure plots posterior means of λ at each recursive forecast origin from 2001M1 to 2019M11 for each model specification. The pronounced elevation during 2008-2010 reflects the Great Recession's elevated volatility. 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 investigates a foundational question in applied macroeconomic forecasting: whether soft information (consumer sentiment) improves inflation and real-activity predictions once one conditions on standard macro aggregates and forward-looking financial prices, and whether adding sentiment refines the model's internal expectation-updating behavior. Using monthly U.S. data over 1985–2019 and a 230-origin expanding-window pseudo out-of-sample design, we estimate hierarchical BVARs under three nested information sets and evaluate performance through both point-forecast accuracy and an expectation-updating diagnostic.

The key empirical findings are as follows. *First*, sentiment contains incremental predictive value for inflation, but primarily at low frequencies (12-month horizons), where it delivers a 4.6% RMSFE reduction relative to a model without sentiment. This is consistent with sentiment capturing information about inflation-expectation anchoring and the persistence of inflation dynamics, which are relevant for medium-run pricing and wage-setting decisions. *Second*, financial variables excel at short-horizon prediction of real activity (9–11% RMSFE improvements at $h = 1, 3$), reflecting their sensitivity to near-term cyclical pressures, but this advantage evaporates at long horizons where secular growth drivers dominate. *Third*, the Coibion-Gorodnichenko revision diagnostic reveals a horizon-dependent pattern of expectation-updating bias: at short horizons ($h = 1$), the baseline model exhibits significant underreaction (positive coefficient 2.41), which sentiment reduces by approximately 65% (to 0.85), though this difference is estimated imprecisely. At long horizons ($h = 12$), all models exhibit overreaction (negative coefficients near -0.56 to -0.64), suggesting that the BVAR's structural estimation of low-frequency movements can lead to trend extrapolation.

Fourth, the hierarchical shrinkage parameter λ adapts automatically to model size: it declines from 0.52 (small model) to 0.42 (medium) to 0.19 (full model), evidence that the prior-selection mechanism is working as intended to maintain comparable regularization intensity across specifications. The Great Recession episode reveals additional time-varying adaptation: λ increases sharply during 2008–2010, allowing the model greater flexibility to capture elevated volatility.

These findings contribute to three research dimensions. For practitioners, they highlight the value of sentiment indices as a source of low-frequency information and

suggest a quantitative benchmark for sentiment's incremental contribution (roughly 0.06 percentage-point RMSFE reduction for 12-month inflation). For methodologists, they demonstrate how hierarchical Bayesian methods can make prior-specification decisions explicit and auditable, avoiding the pitfall of implicit assumptions about regularization intensity. For researchers studying expectation formation, they show that model-based diagnostics (the CG regression) can complement survey-based approaches, offering a systematic way to measure internal consistency in an estimated forecasting system.

The work deliberately avoids claiming more than the data support. Sentiment's effects on the CG regression are estimated with wide confidence intervals, and the causal mechanisms underlying sentiment's predictive value remain unidentified. The paper frames itself as an internal-consistency and transparent-design exercise rather than a definitive verdict on sentiment's economic importance. Future work using real-time data vintages, exploring nonlinear and time-varying relationships, and implementing structural identification will be valuable complements to this foundation.

References

References

Coibion, O., & Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8), 2644–2678.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97(2), 436–451.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions—five years of experience. *Journal of Business & Economic Statistics*, 4(1), 25–38.

Bańbura, M., Giannone, D., & Reichlin, L. (2010). Large Bayesian vector autoregressions. *Journal of Applied Econometrics*, 25(1), 71–92.

Koop, G., & Korobilis, D. (2010). Bayesian multivariate time series methods for empirical macroeconomics. *Foundations and Trends in Econometrics*, 3(4), 267–358.

Kuschnig, N., & Vashold, L. (2021). BVAR: Bayesian vector autoregressions with hierarchical prior selection in R. *Journal of Statistical Software*, 100(14), 1–27. https://doi.org/10.18637/jss.v100.i14

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1), 1–48.

Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2), 147–162.

A Additional figures and robustness

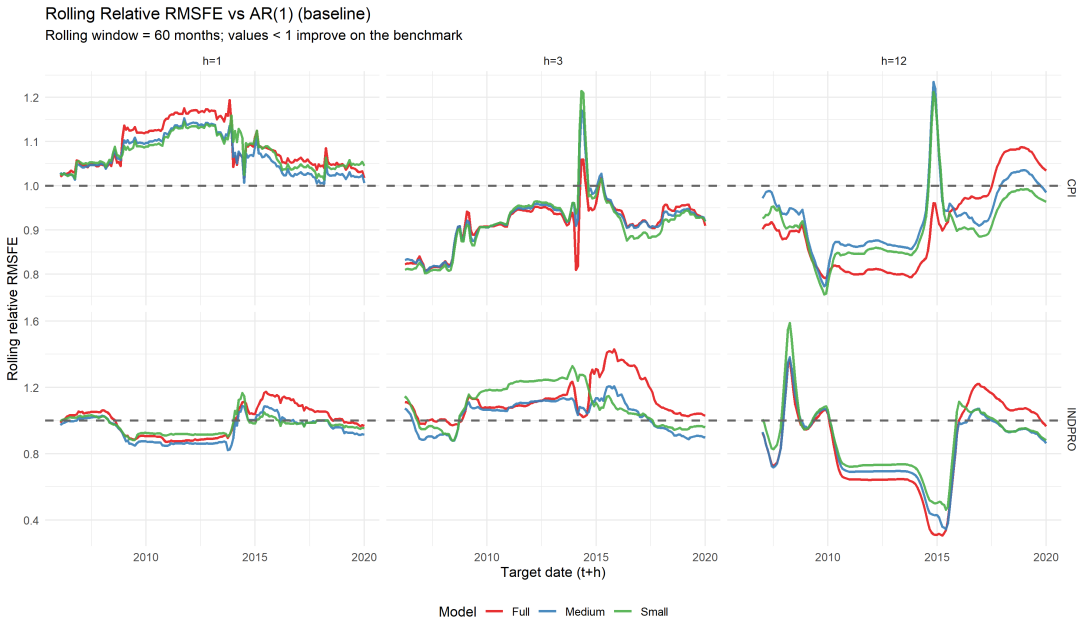

Figure 8: Rolling relative RMSFE versus AR(1) benchmark

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

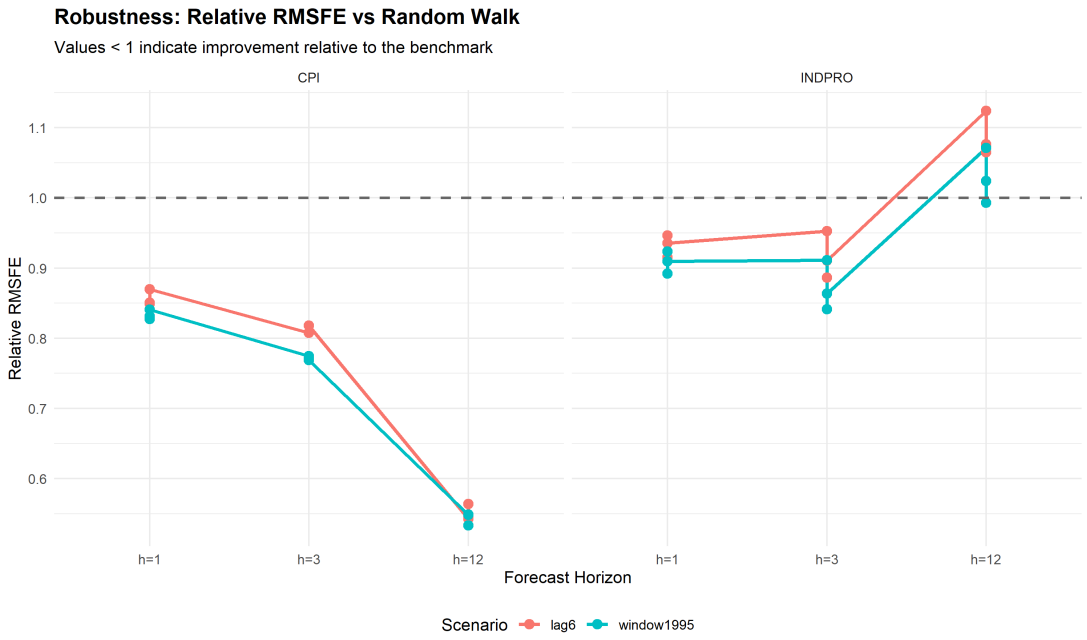

Figure 9: Robustness: relative RMSFE versus no-change benchmark

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.

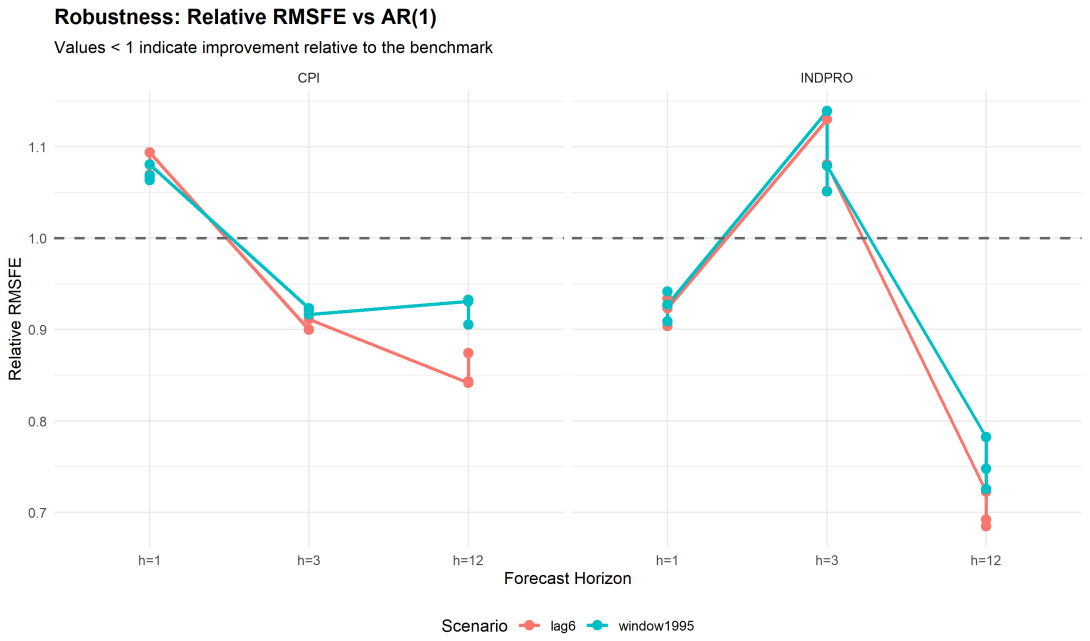

Figure 10: Robustness: relative RMSFE versus AR(1) benchmark

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

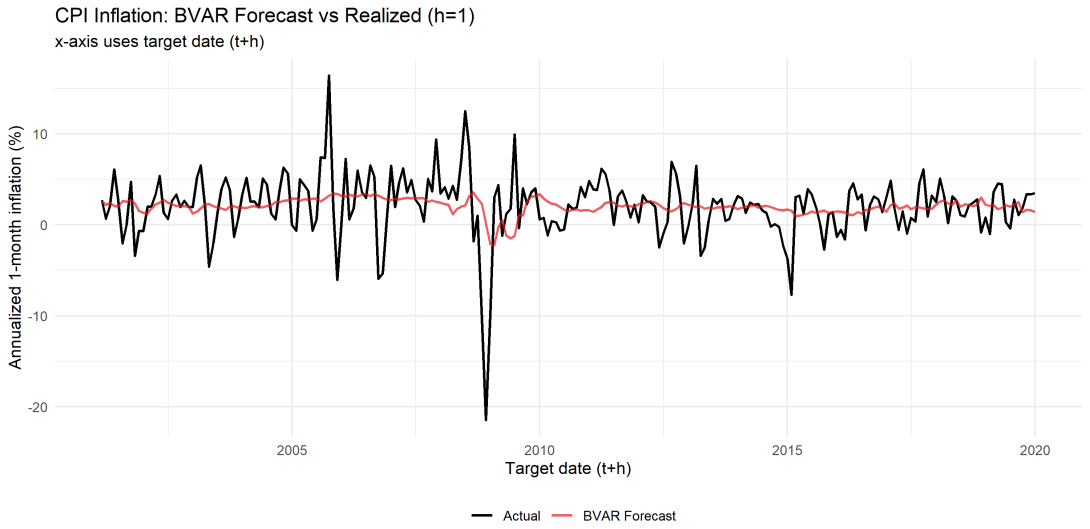

Figure 11: CPI inflation: BVAR forecast versus realized ($h = 1$)

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

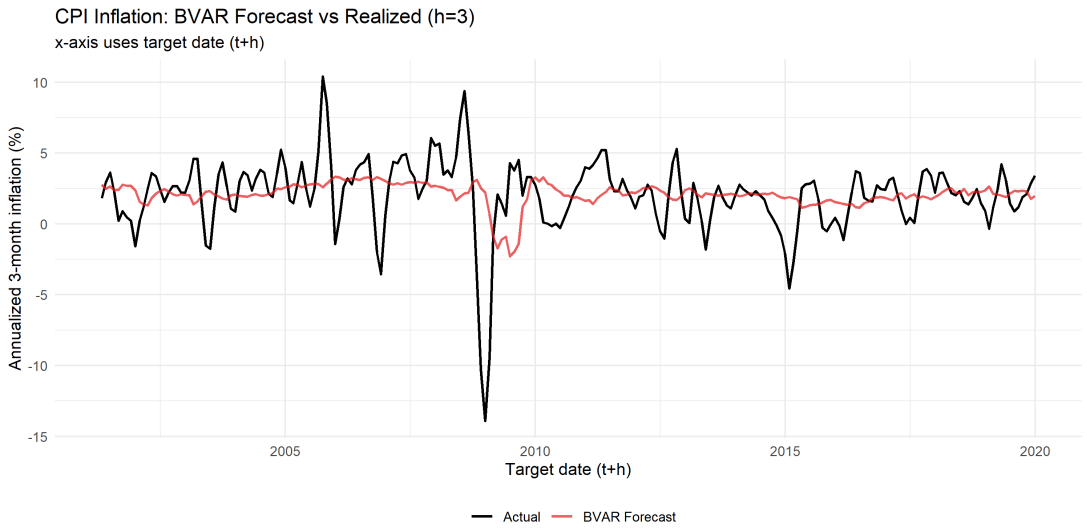

Figure 12: CPI inflation: BVAR forecast versus realized ($h = 3$)

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

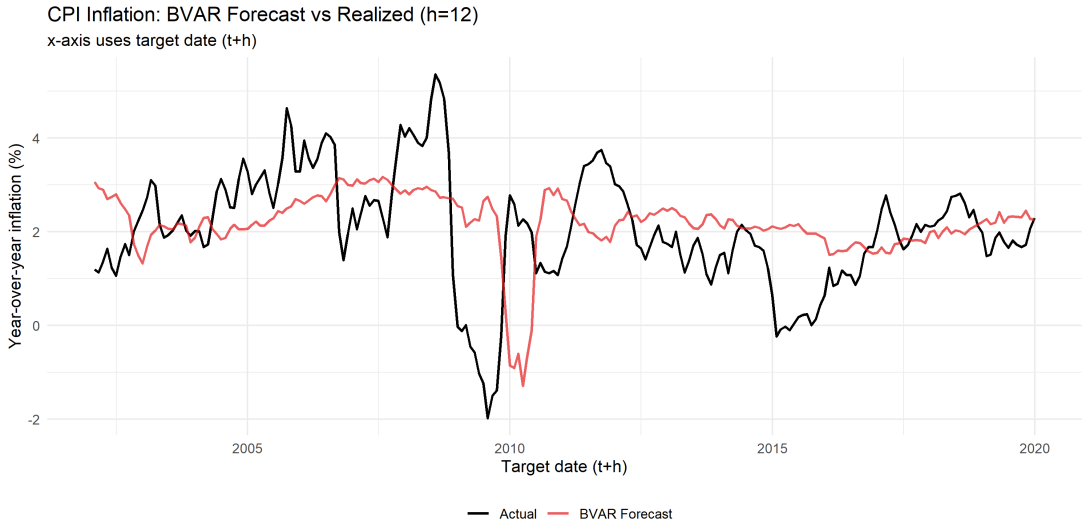

Figure 13: CPI inflation: BVAR forecast versus realized ($h = 12$)

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

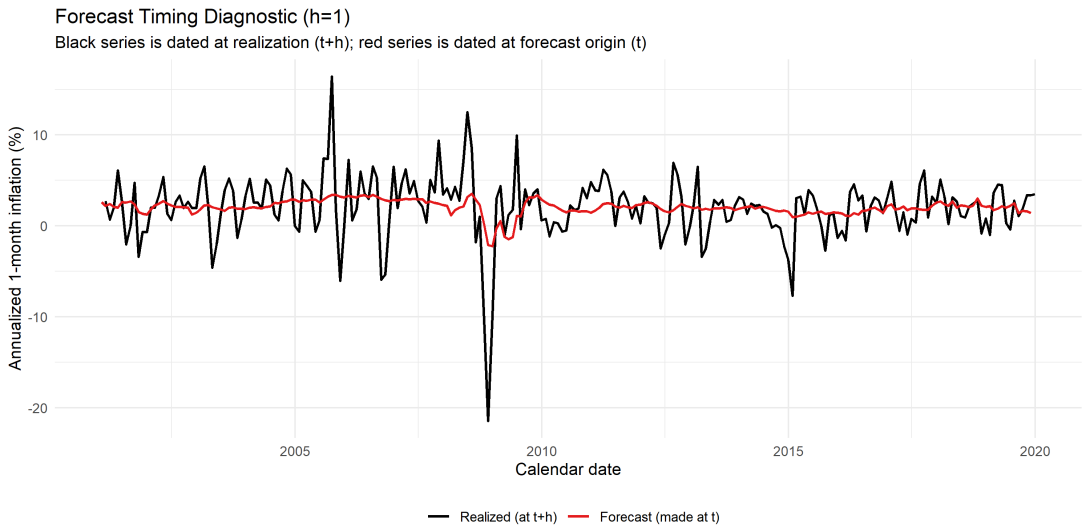

Figure 14: Forecast timing diagnostic ($h = 1$)

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

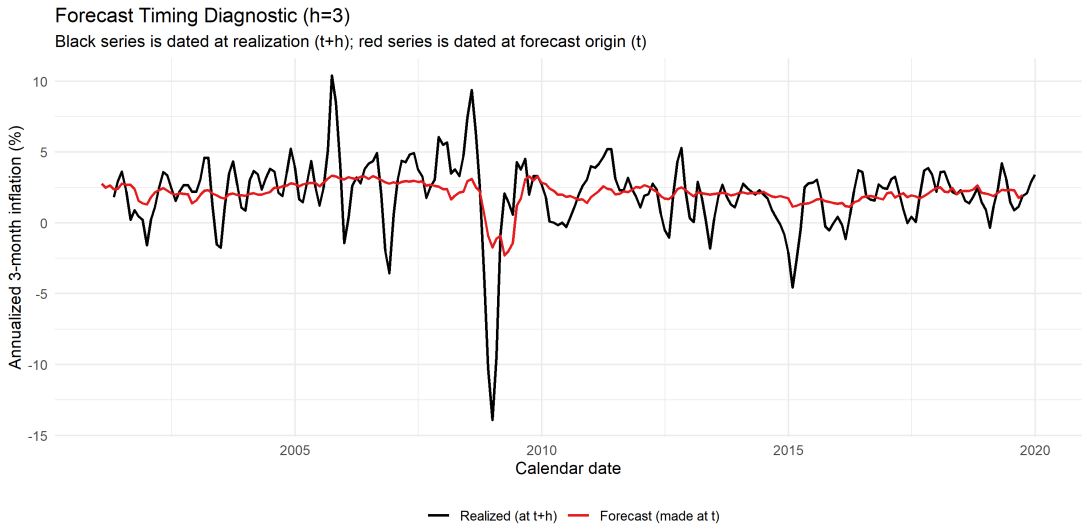

Figure 15: Forecast timing diagnostic ($h = 3$)

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

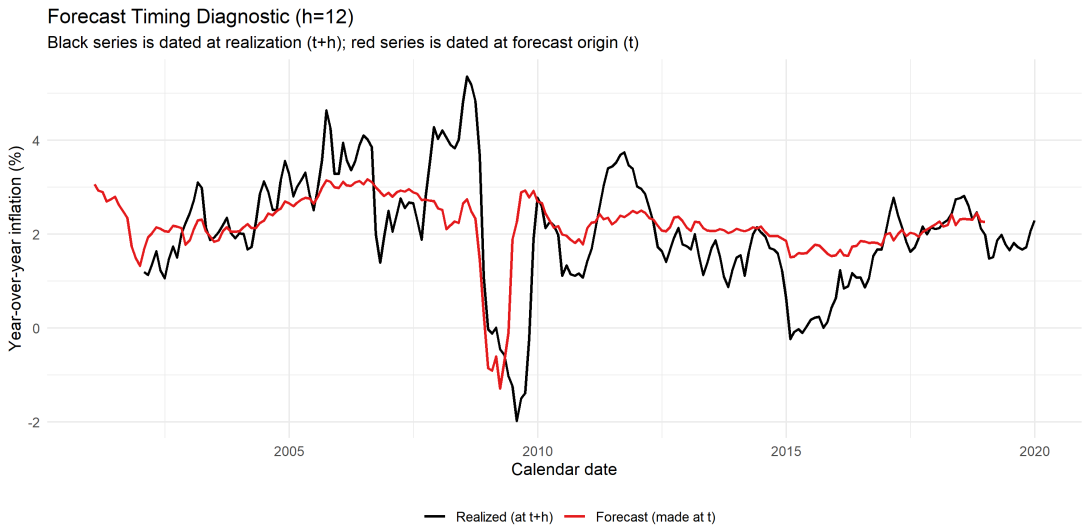

Figure 16: Forecast timing diagnostic ($h = 12$)

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

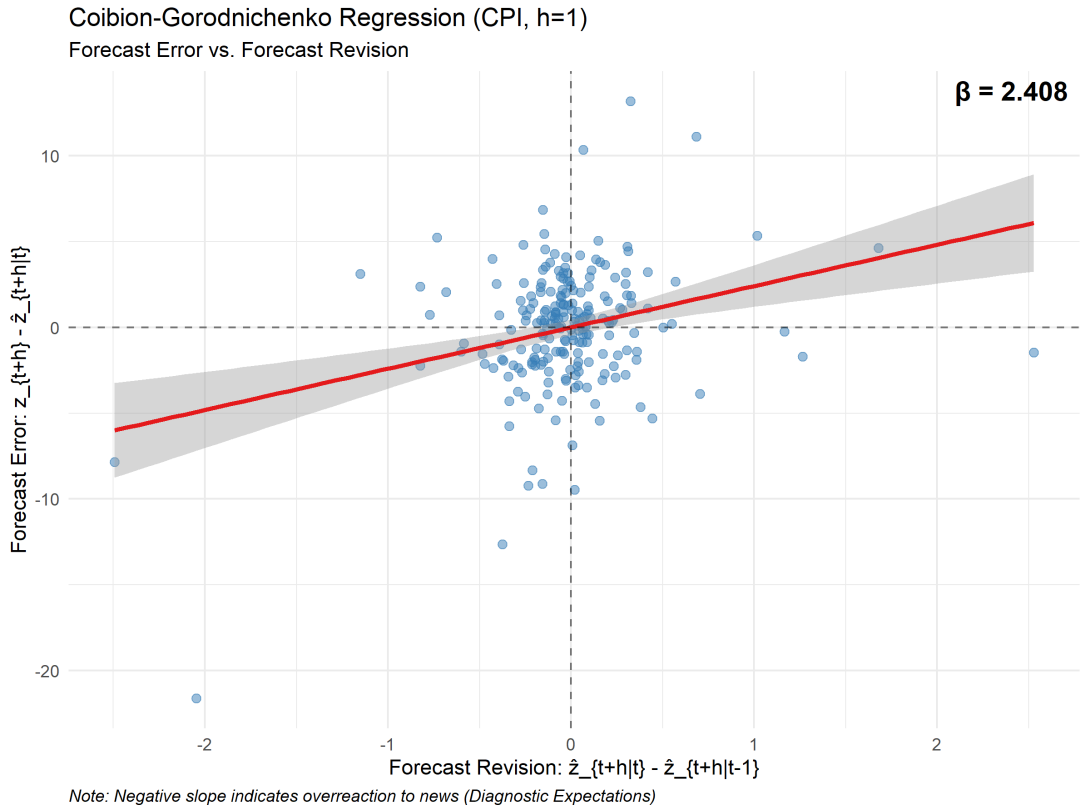

Figure 17: Revision diagnostic scatter: CPI ($h = 1$)

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

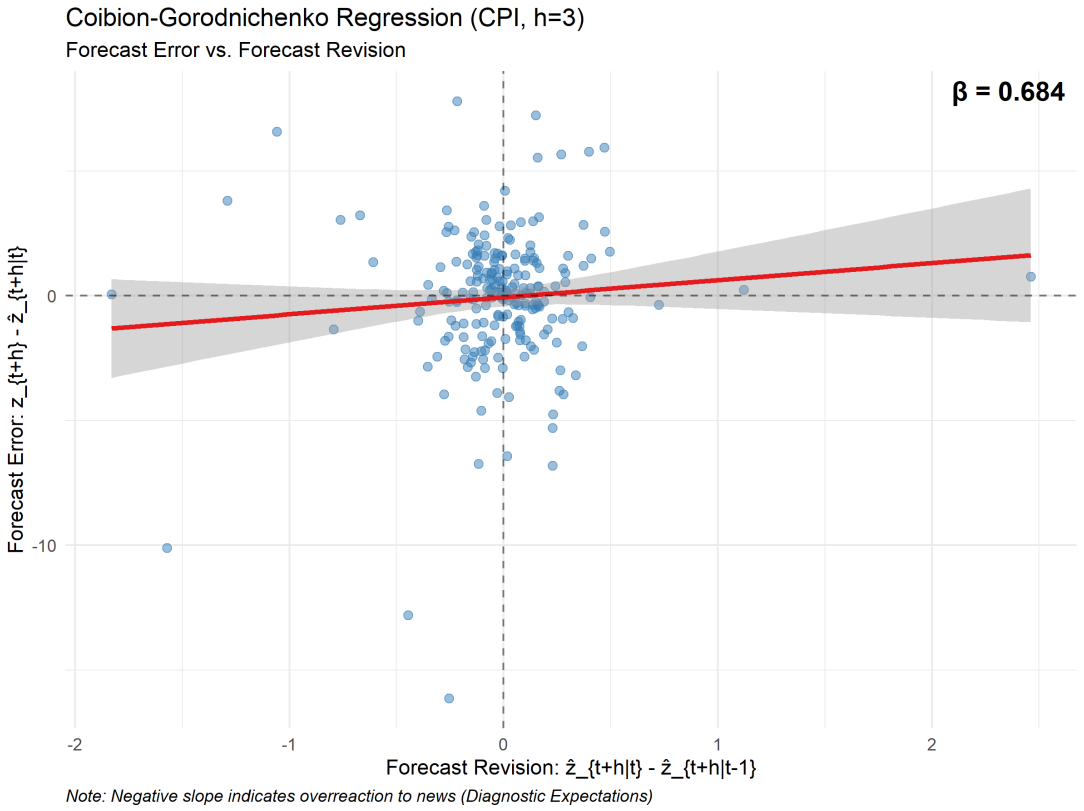

Figure 18: Revision diagnostic scatter: CPI ($h = 3$)

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

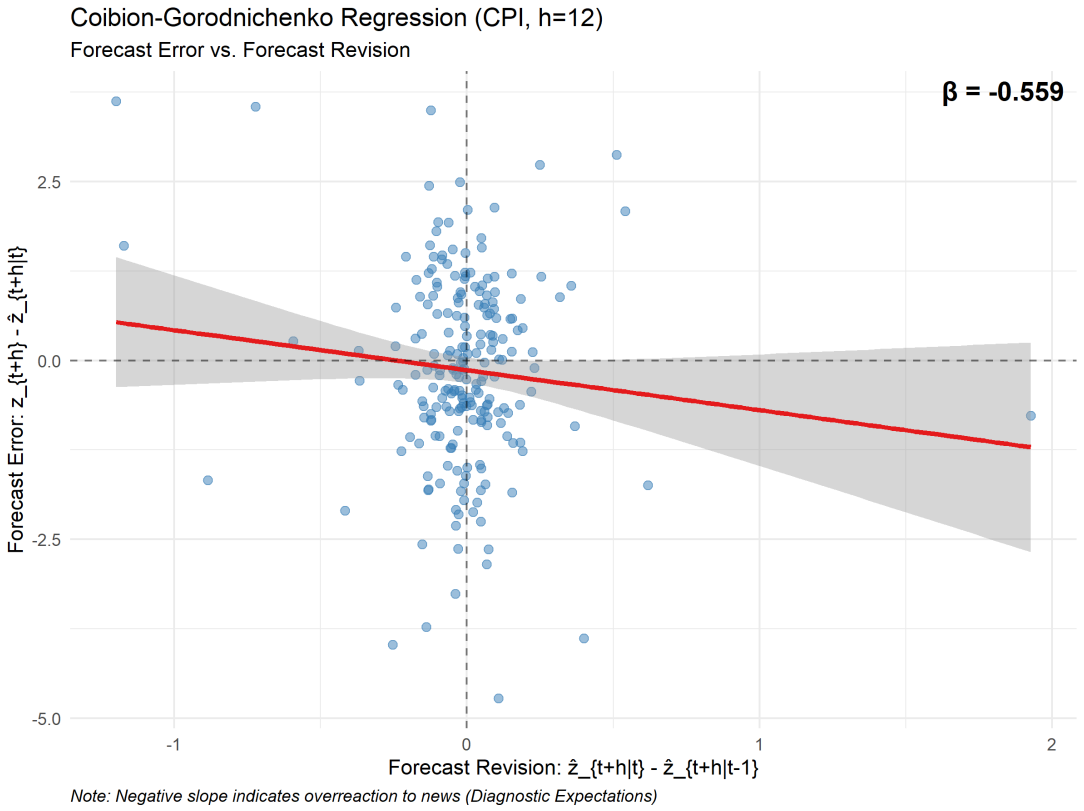

Figure 19: Revision diagnostic scatter: CPI ($h = 12$)

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

Table 4: 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`.