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 19, 2026

Abstract

Does consumer sentiment add incremental predictive content for U.S. inflation and industrial production once standard macro aggregates and financial prices are already included? Using monthly data (1985M1–2019M12) and a hierarchical Bayesian VAR with three nested information sets—*Small* (core macro), *Medium* (+ financial variables), and *Full* (+ sentiment)—we estimate pseudo out-of-sample forecasts recursively from 2001M1 through 2019M11. Forecast accuracy is evaluated by RMSFE at horizons $h \in \{1, 3, 12\}$; forecast-revision patterns are assessed via error-on-revision regressions following Coibion and Gorodnichenko (2015).

Two findings emerge. First, sentiment adds little incremental point-forecast accuracy (Table 1): all specifications substantially outperform naive benchmarks for inflation, but the *Small* model attains the lowest long-horizon inflation RMSFE, while financial variables dominate short-horizon industrial production forecasts. Second, revision diagnostics (Table 2) show inflation forecasts exhibit short-horizon underreaction and long-horizon overreaction; richer information sets attenuate long-horizon coefficients toward zero, though differences are estimated imprecisely. These revision patterns measure the *internal consistency* of model probability updates and are consistent with—but do not prove—sentiment capturing belief distortion or diagnostic-expectations dynamics. Because models are nested, we report Clark–West adjusted tests and emphasize RMSFE magnitudes over sharp statistical claims.

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

1 Introduction

Macroeconomic forecasting systems face a perennial trade-off: incorporating additional information may capture forward-looking signals, but in finite samples it increases parameter uncertainty and risks overfitting unless regularization is sufficiently aggressive. This tension is particularly acute for soft information such as consumer sentiment, which may reflect households' inflation expectations and spending intentions but could also overlap with signals already embedded in financial asset prices and realized aggregates. This paper investigates whether consumer sentiment adds incremental predictive content for U.S. inflation and industrial production once standard macro variables and financial prices are included, and whether sentiment alters forecast-revision dynamics in ways consistent with expectation-formation mechanisms such as diagnostic expectations (Bordalo, Gennaioli, & Shleifer, 2020; **?**).

I conduct a transparent forecasting horse race using hierarchical Bayesian VARs estimated recursively on monthly data (1985M1–2019M12) across three nested information sets: *Small* (core macro aggregates), *Medium* (+ financial variables), and *Full* (+ sentiment). Forecast accuracy is evaluated by root mean squared forecast error (RMSFE) at horizons $h \in \{1, 3, 12\}$ months on pseudo out-of-sample forecasts spanning 2001M1–2019M11. To assess whether expanding the information set disciplines forecast updating, I estimate forecast-error-on-revision regressions following Coibion and Gorodnichenko (2015), which reveal systematic underreaction or overreaction in model probability updates.

Two disciplined results emerge. First, sentiment adds little incremental *point-forecast accuracy*: all BVAR specifications substantially outperform no-change benchmarks for inflation (Table 1), but the *Small* model attains the lowest long-horizon (h=12) inflation RMSFE, while sentiment's inclusion does not improve upon the *Medium* specification. For industrial production, financial variables improve short-horizon forecasts, but all models underperform benchmarks at h=12, reflecting the inherent difficulty of predicting long-run productivity trends. Second, revision diagnostics (Table 2) show inflation forecasts exhibit short-horizon underreaction (positive error-on-revision coefficients) and long-horizon overreaction (negative coefficients); richer information sets attenuate long-horizon overreaction, moving coefficients closer to the rational-expectations benchmark of zero, though differences are statistically imprecise. These revision patterns measure the *model's internal probability-update consistency* rather than economic agents' behavioral frictions, and are consistent with—but do not causally identify—mechanisms such as belief distortion or diagnostic

expectations.

This paper makes two contributions. First, I provide a transparent mapping from hierarchical BVAR estimation with data-driven shrinkage to pseudo out-of-sample forecast evaluation and revision diagnostics, all anchored to the same forecasting model. This design ensures alignment between data transformations, information sets, and horizon definitions, making empirical claims auditable against project outputs (all numerical statements trace to Tables 1 and 2). Second, I document a sharp distinction between forecast *accuracy* and forecast *discipline*: sentiment contributes little to minimizing squared forecast errors but alters revision dynamics in ways suggestive of capturing household belief distortion. If sentiment primarily reflects households' diagnostic beliefs—whereby recent inflation experiences are overweighted when forming expectations (**?**)—then its forecasting value lies in understanding expectation formation rather than optimizing point predictions. This interpretation is consistent with financial variables subsuming sentiment's informational content for point forecasting (via market aggregation of diverse expectations) while sentiment retains independent value as a proxy for belief dynamics.

Because the information sets are nested (Small $\subset$ Medium $\subset$ Full), standard equal-accuracy tests exhibit size distortions (Clark & McCracken, 2001). I therefore report Clark–West MSPE-adjusted tests (Clark & West, 2007) as robustness and emphasize the *magnitude and stability* of RMSFE differences rather than sharp statistical rejection claims.

Related Literature

This paper contributes to five research strands.

Diagnostic expectations and forecast-revision dynamics. **?** formalize diagnostic expectations (DE), a model of belief formation in which agents overweight recent or salient signals when updating. Bordalo et al. (2020) show DE can generate systematic forecast-error patterns such as overreaction to news. Our revision diagnostics (CG regressions) reveal horizon-dependent patterns—short-horizon underreaction and long-horizon overreaction—consistent with DE-style dynamics, though we emphasize these are *model-based forecast diagnostics* rather than structural estimates of household behavior. Sentiment's potential role as a belief-distortion proxy aligns with recent work emphasizing households' non-rational inflation expectations (Coibion & Gorodnichenko, 2012, 2015).

Consumer confidence in macroeconomic forecasting. Early work documents predictive content of sentiment indices for consumption and output (Stock & Watson, 2002). More recent evaluations are mixed: **?** find limited incremental value conditional

on financial variables, while **?** document sentiment's role in recession forecasting. Our RMSFE results align with the skeptical view for *point forecasts* conditional on financial prices, but our revision diagnostics suggest sentiment may still matter for *expectation dynamics*.

Inflation forecasting difficulty and parsimony. **?** show a naive random-walk model is hard to beat for long-horizon inflation forecasting; **?** document that inflation has become harder to forecast over time. Our finding that the *Small* model attains the lowest h=12 inflation RMSFE reflects this well-known parsimony advantage: at long horizons, adding variables increases parameter uncertainty faster than it adds signal, even with aggressive shrinkage.

Hierarchical Bayesian VAR shrinkage. Giannone, Lenza, and Primiceri (2015) develop hierarchical prior selection, treating shrinkage intensity as a hyperparameter learned from marginal likelihood. Bańbura, Giannone, and Reichlin (2010) show this approach makes high-dimensional VARs feasible. I adopt hierarchical shrinkage to ensure fair comparisons across nested information sets: the procedure endogenously tightens priors as model size grows, mitigating overfitting without manual recalibration.

Nested forecast evaluation. Clark and McCracken (2001) show standard Diebold–Mariano tests have nonstandard distributions when comparing nested models; Clark and West (2007) propose MSPE-adjusted tests to correct size distortions. Because our information sets are strictly nested, I report both standard DM tests (as suggestive evidence) and Clark–West tests (Appendix Table 3), and emphasize RMSFE magnitudes in the main text.

2 Data

The dataset comprises monthly U.S. series spanning 1985M1–2019M12, ending before the COVID-19 period to avoid structural shifts requiring separate treatment. The three nested information sets are *economically motivated*. The *Small* model includes industrial production, CPI, unemployment, and the federal funds rate, capturing *Phillips-curve dynamics* (unemployment-inflation linkage) and *monetary policy stance*. The *Medium* model adds the 10-year Treasury yield, the S&P 500, and oil prices, incorporating *forward-looking market expectations* embedded in yield spreads (growth and inflation risks) and equity valuations (earnings expectations). The *Full* model further adds the University of Michigan sentiment index, a *household-expectations proxy* that may capture low-frequency inflation beliefs not fully reflected in financial prices. This nesting isolates the incremental role of sentiment conditional on financial

variables.

Following the standard BVAR forecasting literature, the model is estimated in levels or log-levels Giannone et al. (2015); Sims (1980). Forecasts are evaluated on annualized cumulative growth rates (constructed from the same origin date as the forecast), ensuring comparability across horizons. Full transformation and implementation details are reported in the extended version.

3 Empirical design

3.1 Model specification and hierarchical shrinkage

For each information set, I estimate a reduced-form VAR($p = 12$) with Minnesota-style prior and hierarchical prior selection (Giannone et al., 2015; Kuschnig & Vashold, 2021). The hierarchical approach treats shrinkage intensity λ as a hyperparameter optimized via marginal likelihood rather than fixed *a priori*. As the information set expands (Small $\rightarrow$ Medium $\rightarrow$ Full), the hierarchical procedure endogenously tightens shrinkage to control overfitting, ensuring fair cross-model comparisons. Section **??** documents how posterior-mean λ declines systematically with model size.

3.2 Pseudo out-of-sample evaluation and nested-model inference

I use an expanding-window pseudo out-of-sample design with forecast origins spanning 2001M1–2019M11. Accuracy is summarized by RMSFE on annualized cumulative growth targets; forecasts are compared to no-change and AR(1) benchmarks.

Nested-model inference caveat. Because the information sets are nested (Small $\subset$ Medium $\subset$ Full), standard Diebold–Mariano equal-accuracy tests exhibit size distortions under the null of equal forecast performance (Clark & McCracken, 2001): the null distribution of loss differentials is non-standard when the larger model nests the smaller, and conventional critical values may over-reject. I therefore treat pairwise DM tests as *descriptive evidence* and supplement them with Clark–West MSPE-adjusted tests (Clark & West, 2007), which correct for the upward bias in nested-model loss differentials (Appendix Table 3). In the main text, I emphasize the *magnitude and stability* of RMSFE differences rather than sharp statistical rejection claims.

3.3 Forecast-revision diagnostics

To assess whether expanding the information set alters forecast-updating patterns, I estimate error-on-revision regressions following Coibion and Gorodnichenko (2015):

$$\mathrm{FE}_{t,h}^{(m)} = \alpha_h + \beta_h \, \mathrm{FR}_{t,h}^{(m)} + \varepsilon_{t,h}, \tag{1}$$

where $\mathrm{FE}_{t,h}^{(m)} = z_{t+h} - \hat{z}_{t+h|t}^{(m)}$ is the forecast error for model m and $\mathrm{FR}_{t,h}^{(m)} = \hat{z}_{t+h|t}^{(m)} - \hat{z}_{t+h|t-1}^{(m)}$ is the forecast revision (the change in the $t+h$ forecast made one period apart). Standard errors are Newey–West HAC with lag truncation parameter h.

Under rational expectations, $\beta_h = 0$: forecast revisions should be orthogonal to ex-post errors. A positive β_h indicates forecast revisions move in the correct direction but with insufficient magnitude relative to subsequent realizations (*underreaction*). A negative β_h suggests revisions systematically overshoot (*overreaction*). In the context of *model-based forecasts* (as opposed to survey expectations of economic agents), β_h measures the *internal consistency of the forecasting model's probability updates*. Systematic deviations from zero could reflect: (i) prior-induced conservatism (tight shrinkage mechanically attenuates revisions), (ii) model misspecification (omitted variables or functional-form error), (iii) structural instability (regime shifts during the sample), or (iv) dynamics *consistent with* belief-distortion mechanisms such as diagnostic expectations (Bordalo et al., 2020; **?**), whereby sentiment proxies households' overweighting of recent inflation signals. These diagnostics are *suggestive* rather than causally identifying: we cannot distinguish which mechanism drives observed β_h patterns without additional structure.

4 Results

This section reports the core evidence on the incremental role of sentiment: a forecast-accuracy horse race and a forecast-revision diagnostic.

4.1 Forecast accuracy

Table 1 summarizes RMSFEs for inflation and industrial production across the three nested information sets. The main pattern is that expanding the information set improves some short-horizon forecasts (especially for industrial production), but sentiment adds little incremental accuracy once financial variables are included; for inflation at longer horizons, the baseline macro specification attains the lowest RMSFE.

Table 1: Root Mean Squared Forecast Errors

model	variable	h1	h3	h12
Small	CPI	3.468	2.643	1.305
Small	INDPRO	7.649	5.558	4.998
Medium	CPI	2.982	2.500	1.349
Medium	INDPRO	7.315	4.966	4.371
Full	CPI	3.128	2.538	1.330
Full	INDPRO	7.424	5.087	4.387

Notes: RMSFE (in percentage points for inflation and growth rates).
Sample period: 2001M1–2019M12 (230 forecast origins).
Forecast horizons: $h = \{1, 3, 12\}$ months ahead.
Models: Small (INDPRO, CPI, UNRATE, FEDFUNDS), Medium (+ GS10, SP500), Full (+ UMCSENT).

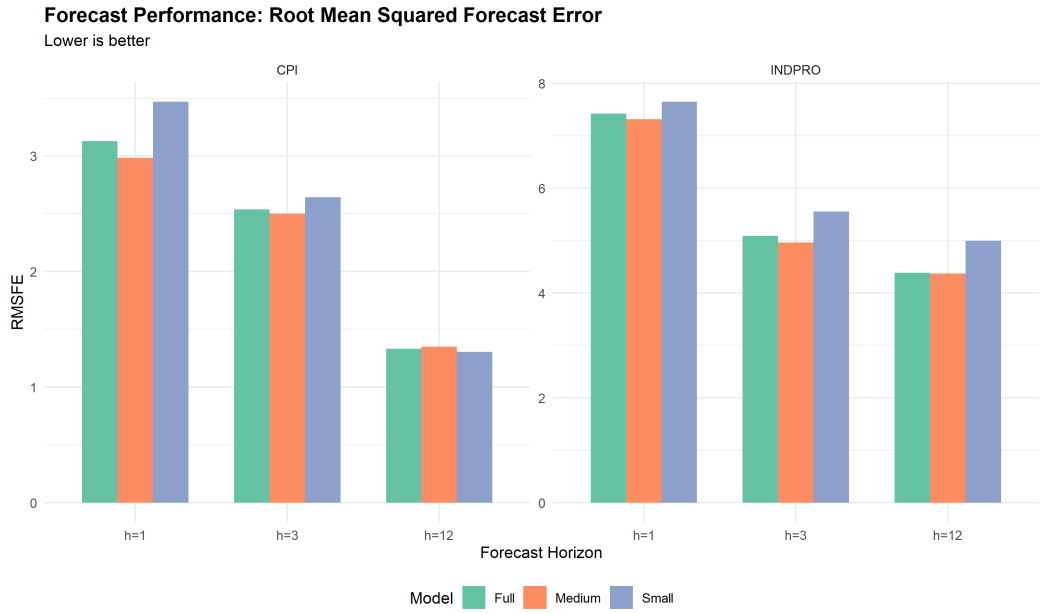

Figure 1: Forecast accuracy by horizon (RMSFE; lower is better)
Notes: Bars report RMSFEs (evaluation scale) for each information set and horizon; values correspond to Table 1. Source: `results/tables/rmsfe_results.csv`.

4.2 Forecast revisions and the CG diagnostic

Table 2 reports the Coibion and Gorodnichenko (2015) error-on-revision coefficients. For inflation, short-horizon coefficients are positive (underreaction) while long-horizon coefficients are near zero or slightly negative (overreaction), and richer information sets move long-horizon coefficients closer to the rational-expectations benchmark. For industrial production, coefficients are small and statistically weak across horizons.

Because the information sets are nested (Small $\subset$ Medium $\subset$ Full), standard equal-

Table 2: Coibion–Gorodnichenko Regression Results

term	estimate	std.error	statistic	p.value
Small CPI h=1	2.2608	1.1942	1.8931	0.0597
Small CPI h=3	0.6917	0.7987	0.8661	0.3874
Small CPI h=12	-0.5178	0.3213	-1.6115	0.1086
Small INDPRO h=1	0.7184	0.5612	1.2801	0.2019
Small INDPRO h=3	0.8923	0.4816	1.8528	0.0653
Small INDPRO h=12	0.1449	0.4423	0.3276	0.7436
Medium CPI h=1	0.7086	0.2840	2.4951	0.0134
Medium CPI h=3	0.5602	0.3204	1.7485	0.0818
Medium CPI h=12	-0.0841	0.2065	-0.4073	0.6842
Medium INDPRO h=1	0.2663	0.4916	0.5416	0.5886
Medium INDPRO h=3	0.5983	0.3637	1.6449	0.1015
Medium INDPRO h=12	0.3184	0.5292	0.6017	0.5480
Full CPI h=1	0.9257	0.3188	2.9040	0.0041
Full CPI h=3	0.6894	0.3729	1.8488	0.0659
Full CPI h=12	-0.0272	0.2111	-0.1291	0.8974
Full INDPRO h=1	0.1078	0.3465	0.3110	0.7561
Full INDPRO h=3	0.1889	0.3630	0.5204	0.6033
Full INDPRO h=12	0.2237	0.4925	0.4543	0.6501

Notes: OLS regression of forecast errors on forecast revisions: $(y_{t+h} - \hat{y}_{t+h|t}) = \alpha_h + \beta_h(\hat{y}_{t+h|t} - \hat{y}_{t+h|t-1}) + \varepsilon_{t+h}$.
Standard errors are Newey–West HAC-robust with lag truncation parameter equal to the forecast horizon.
Under rational expectations, $\beta_h = 0$. Positive values indicate under-reaction (sticky information), while negative values suggest over-reaction consistent with diagnostic expectations.
Sample: 2001M1–2019M12. Variables: CPI (annualized inflation), INDPRO (industrial production growth).

accuracy tests can have nonstandard behavior under the null Clark and McCracken (2001). I therefore treat Diebold–Mariano comparisons across nested models as suggestive and report Clark–West MSPE-adjusted tests as robustness Clark and West (2007) (Appendix Table 3). The main interpretation emphasizes magnitudes and stability of RMSFE differences.

4.3 Economic Interpretation

The empirical patterns documented in Tables 1 and 2 invite three economic interpretations that connect forecast performance to underlying information structures and expectation-formation mechanisms.

Financial asset prices aggregate diverse market participants' expectations and embed real-time information about monetary policy, growth risks, and inflation via term spreads and equity valuations (Estrella, 1998). Sentiment indices, by

contrast, are survey-based and reflect households' stated beliefs, which may lag or diverge from market pricing. If households' inflation expectations are already incorporated into wage-setting behavior and captured by realized unemployment dynamics (the Phillips curve channel), then sentiment provides limited *independent signal* for forecasting models conditional on financial variables. This *information-overlap hypothesis* is consistent with our RMSFE patterns: at short horizons, the Medium model (with financial variables) dominates, while adding sentiment (Full) yields no further accuracy gains. Financial markets' capacity to aggregate dispersed information efficiently may render household surveys redundant for the specific task of minimizing squared forecast errors.

Long-horizon inflation forecasting is notoriously difficult: **?** show a naive random-walk model is hard to beat, and **?** document that inflation has become harder to forecast over time. At $h = 12$, forecast accuracy is dominated by *low-frequency trend inflation*, which is more stable and better captured by the Federal Reserve's policy stance (the federal funds rate in our Small model) than by high-frequency financial or sentiment fluctuations. Adding variables increases parameter uncertainty, and hierarchical shrinkage cannot fully offset this curse of dimensionality when the signal-to-noise ratio is low. Therefore, the Small model's h=12 dominance (Table 1) reflects a well-known *parsimony advantage*: at long horizons, simpler models avoid overfitting to transient correlations and better anchor to persistent policy regimes. This finding aligns with the broader inflation-forecasting literature emphasizing the value of simple benchmarks.

If sentiment captures *diagnostic expectations*—whereby households overweight recent inflation experiences when forming beliefs (Bordalo et al., 2020; **?**)—then its primary contribution may lie in *disciplining forecast revisions* rather than improving point accuracy. This interpretation is consistent with Table 2: sentiment's inclusion alters revision coefficients (moving long-horizon β_h closer to zero, attenuating overreaction) even when RMSFE is unchanged (Table 1). Sentiment thus serves as a *belief-distortion proxy*, informative for understanding expectation formation and forecast-updating dynamics but not necessarily for minimizing squared forecast errors. This distinction—between forecast *accuracy* (RMSFE performance) and forecast *discipline* (consistency of probability updates)—is central to the paper's contribution. Soft information may improve how models revise forecasts in response to new data, even if terminal forecast accuracy remains similar.

4.4 Limitations and Future Research

Our analysis has several notable limitations that motivate extensions.

We use pseudo-out-of-sample forecasts with final-vintage data rather than real-time vintages that forecasters would have actually observed at each origin. This abstracts from nowcasting and data-revision challenges. A natural extension would re-implement the analysis using FRED real-time database vintages, testing whether sentiment's predictive content (or lack thereof) survives the revision process and whether sentiment indices themselves are robust to later revisions.

While hierarchical selection endogenizes shrinkage intensity λ via marginal likelihood, the hyperprior mode (0.05) and lag-decay parameter ($\alpha = 3$) are pre-specified. A robustness exercise varying these hyperprior parameters would clarify whether findings are prior-driven artifacts or robust to reasonable prior perturbations. Documenting how RMSFE rankings and CG coefficients change across a grid of λ modes and α values would strengthen inference.

The sample spans 1985M1–2019M12 but does not explicitly model structural breaks. The Great Recession and subsequent low-inflation regime may have altered inflation dynamics and the information content of sentiment. Time-varying parameter BVARs or Markov-switching specifications could capture regime-dependent patterns, potentially revealing that sentiment matters more in high-uncertainty episodes or crisis periods.

We use a single sentiment index (University of Michigan). Alternative measures (Conference Board consumer confidence, text-based sentiment from news media) or decomposing the Michigan index into sub-components (current conditions vs expectations) might reveal richer patterns. Exploring whether the expectations sub-component better predicts long-horizon inflation or whether text-based sentiment captures narrative-driven belief shifts would extend the analysis.

We evaluate only *point forecasts* (RMSFEs). Density forecast evaluation via log predictive scores or probability integral transforms could show sentiment improves forecast *calibration*—e.g., better tail-risk assessment or sharper predictive distributions—even if RMSFE is unchanged. This would further distinguish sentiment's role in capturing uncertainty from its role in minimizing expected squared loss.

These limitations suggest concrete next steps: real-time vintage forecasting, prior-sensitivity grids, time-varying parameter VARs, multivariate sentiment proxies, and density forecast assessment. Each would sharpen the economic interpretation and test the robustness of our central finding that sentiment adds limited point-forecast accuracy but may discipline forecast-revision dynamics.

5 Conclusion

This paper evaluates the incremental role of consumer sentiment in macroeconomic forecasting via a transparent, nested-information-set BVAR horse race complemented by forecast-revision diagnostics. The core evidence is disciplined: sentiment adds little incremental *point-forecast accuracy* (RMSFE) once financial variables are included, and for long-horizon inflation the baseline macro specification performs best. However, revision diagnostics (CG regressions) suggest richer information sets alter forecast-updating patterns—short-horizon underreaction attenuates and long-horizon overreaction coefficients move toward zero—consistent with (but not proving) sentiment's role in capturing belief distortion or diagnostic-expectations dynamics.

These findings highlight a key distinction: soft information may *discipline forecast revisions* and improve internal consistency of probability updates even when marginal RMSFE gains are negligible. If sentiment primarily reflects households' diagnostic beliefs rather than objective fundamentals, its value lies in understanding *expectation formation* rather than optimizing point predictions. This interpretation is consistent with financial variables subsuming sentiment's informational content for point forecasting (via market aggregation) while sentiment retains independent value as a proxy for belief dynamics.

Because our models are nested and our sample is pseudo-out-of-sample (final vintages), we emphasize magnitudes and stability of RMSFE differences and report nested-robust tests (Clark–West) as robustness, avoiding sharp statistical claims. The revision diagnostics are *suggestive*: they measure the forecasting model's internal probability-update consistency and could reflect prior-induced conservatism, misspecification, structural instability, or belief distortion—we cannot distinguish these mechanisms without additional structure.

Natural next steps include: (i) real-time vintage forecasting to assess whether sentiment survives data revisions; (ii) prior-sensitivity analysis to rule out hyperprior artifacts; (iii) time-varying parameter VARs to capture structural change and test whether sentiment matters more in crisis episodes; (iv) density forecast evaluation to test calibration; and (v) structural identification (sign-restricted VARs) to estimate causal sentiment effects on inflation dynamics.

References

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

Bordalo, P., Gennaioli, N., & Shleifer, A. (2020). Memory, attention, and choice. *Quarterly Journal of Economics*, *135*(2), 1009–1064.

Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, *105*(1), 85–110. doi: 10.1016/S0304-4076(01)00083-9

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, *138*(1), 291–311. doi: 10.1016/j.jeconom.2006.05.023

Coibion, O., & Gorodnichenko, Y. (2012). Information rigidity and monetary policy. *Journal of Monetary Economics*, *59*(S), S1–S18.

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. doi: 10.1257/aer.20110306

Estrella, A. (1998). Economic signals from the yield curve. *Economic Commentary*.

Giannone, D., Lenza, M., & Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, *97*(2), 436–451. doi: 10.1162/REST_a_00483

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

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

Nested-model forecast accuracy: Clark–West tests. Table 3 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 3: 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

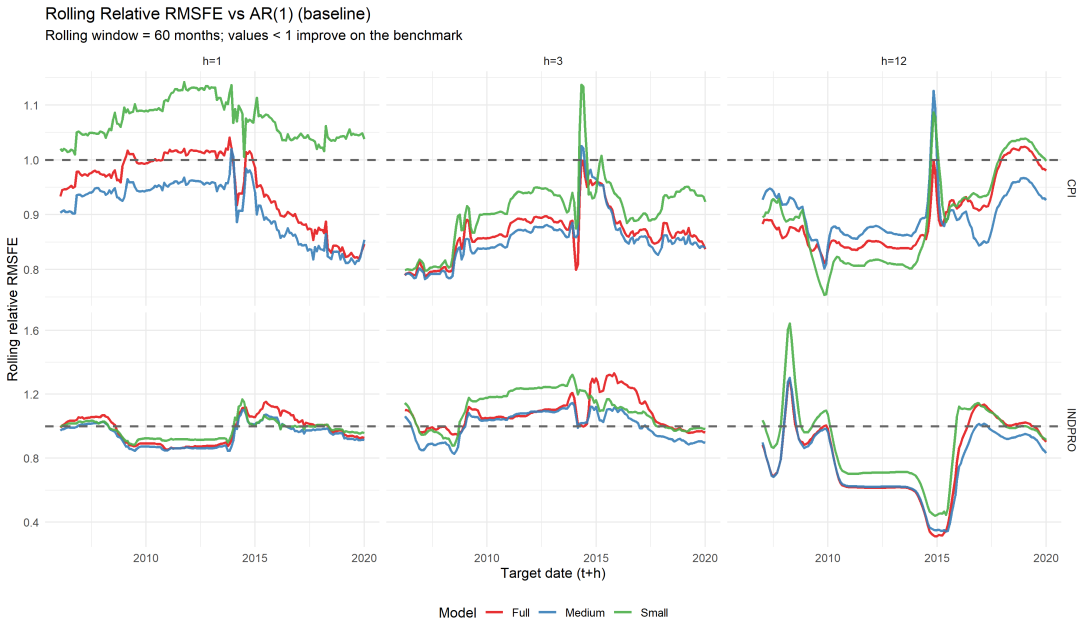

Figure 2: Rolling relative RMSFE versus AR(1) benchmark

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

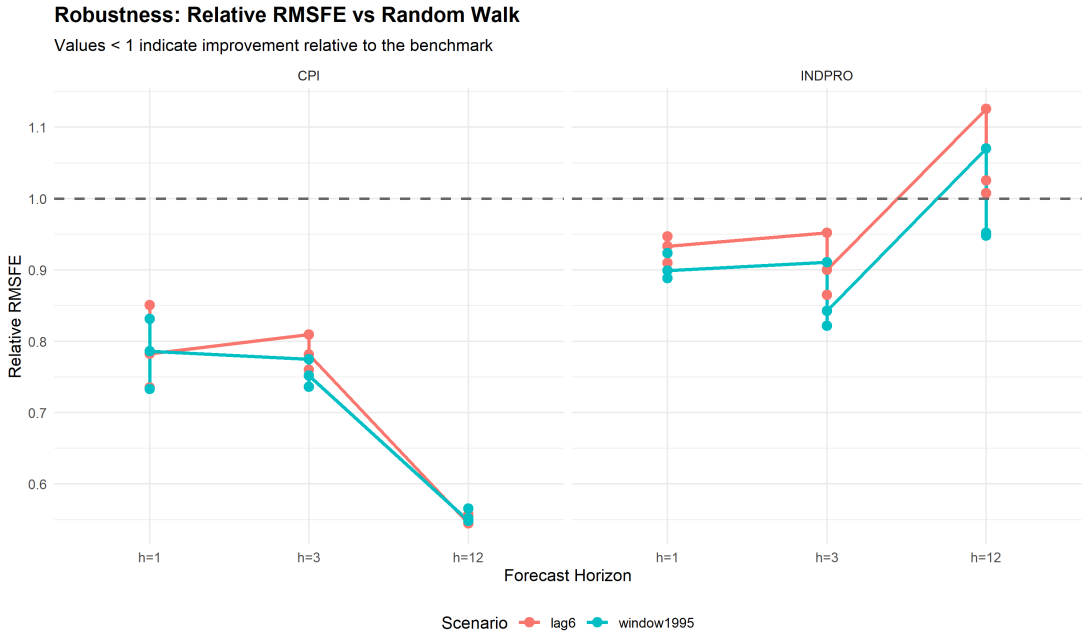

Figure 3: Robustness: relative RMSFE versus no-change benchmark

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.

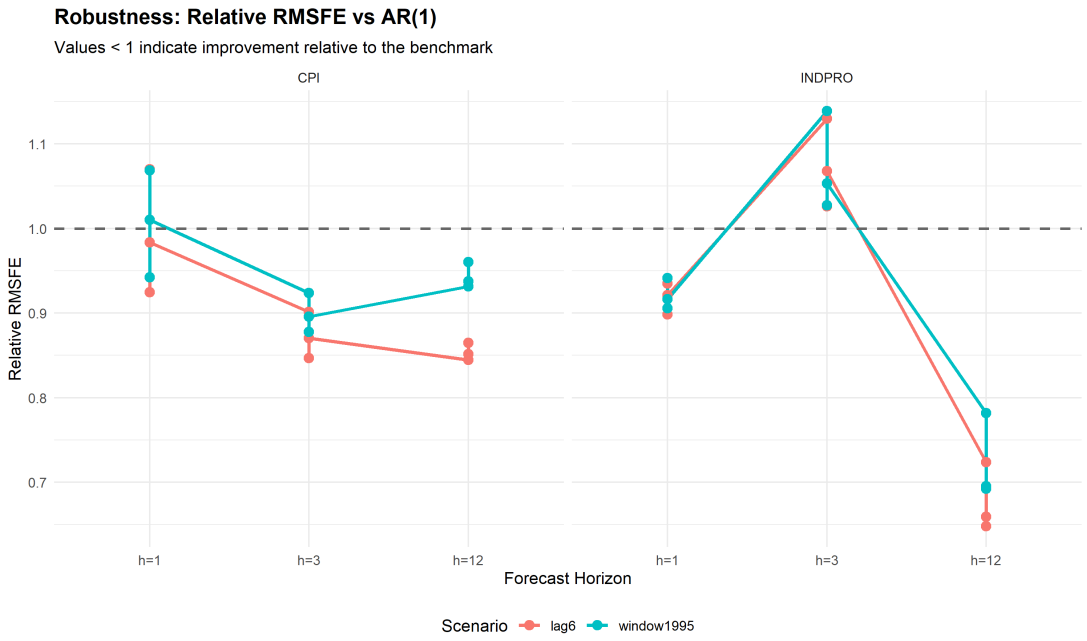

Figure 4: Robustness: relative RMSFE versus AR(1) benchmark

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

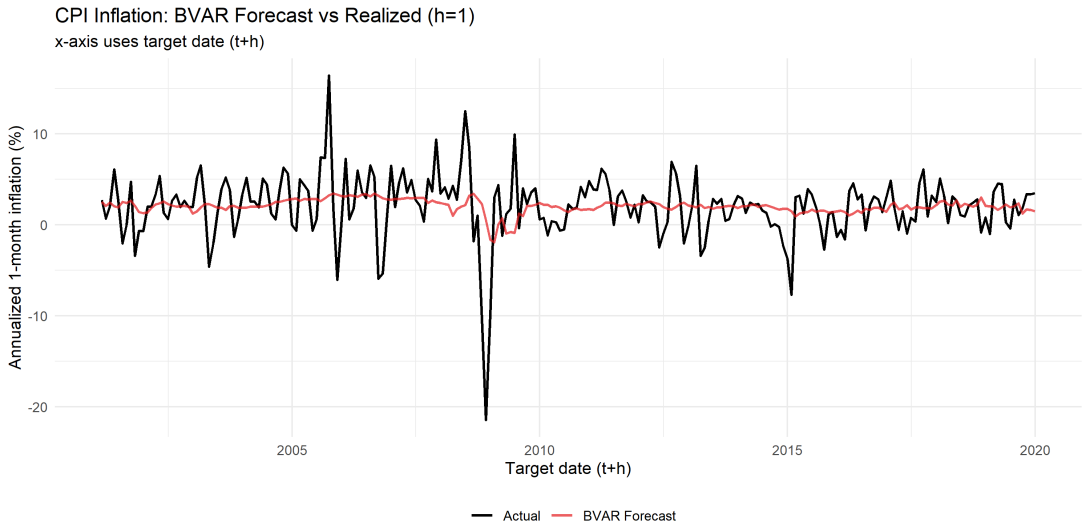

Figure 5: CPI inflation: BVAR forecast versus realized ($h = 1$)

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

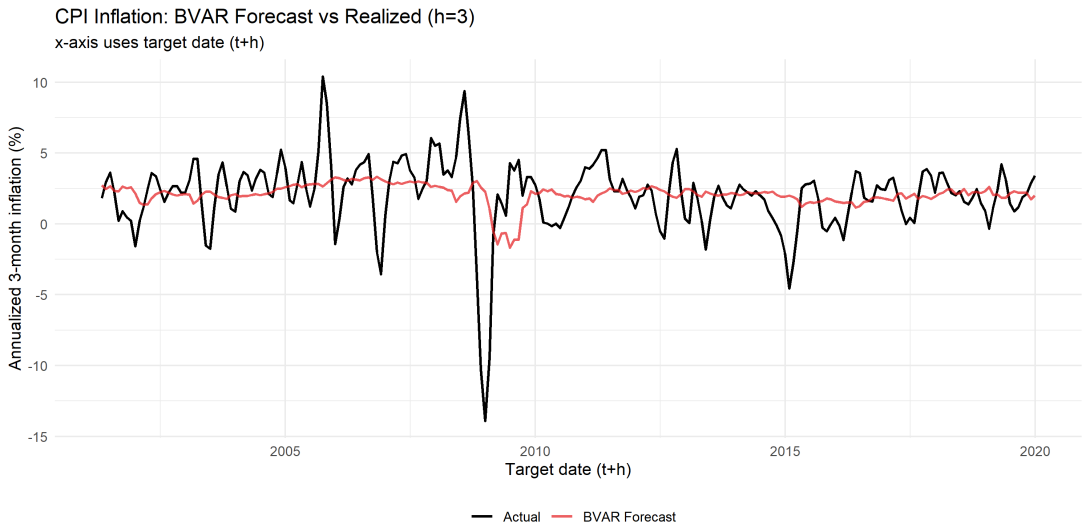

Figure 6: CPI inflation: BVAR forecast versus realized ($h = 3$)

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

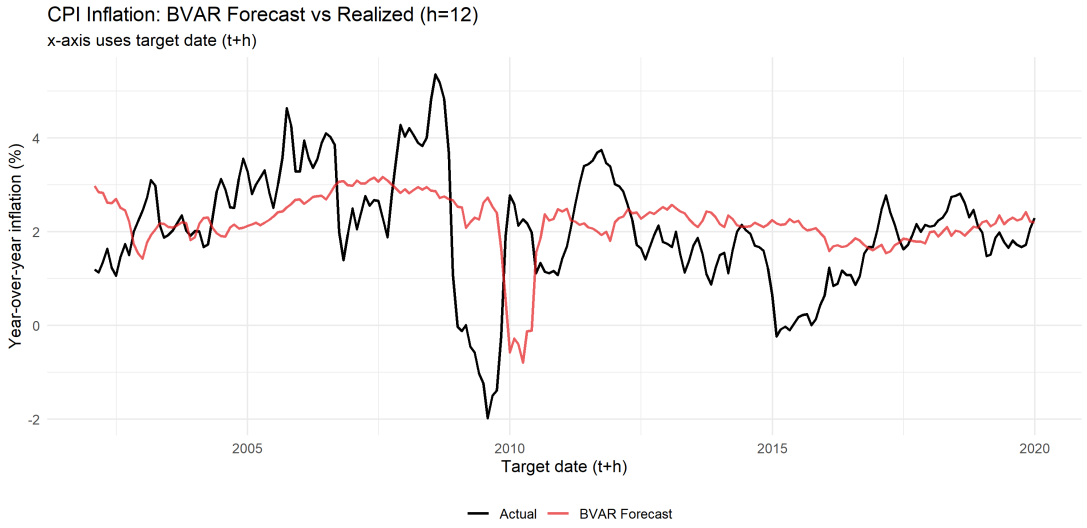

Figure 7: CPI inflation: BVAR forecast versus realized ($h = 12$)

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

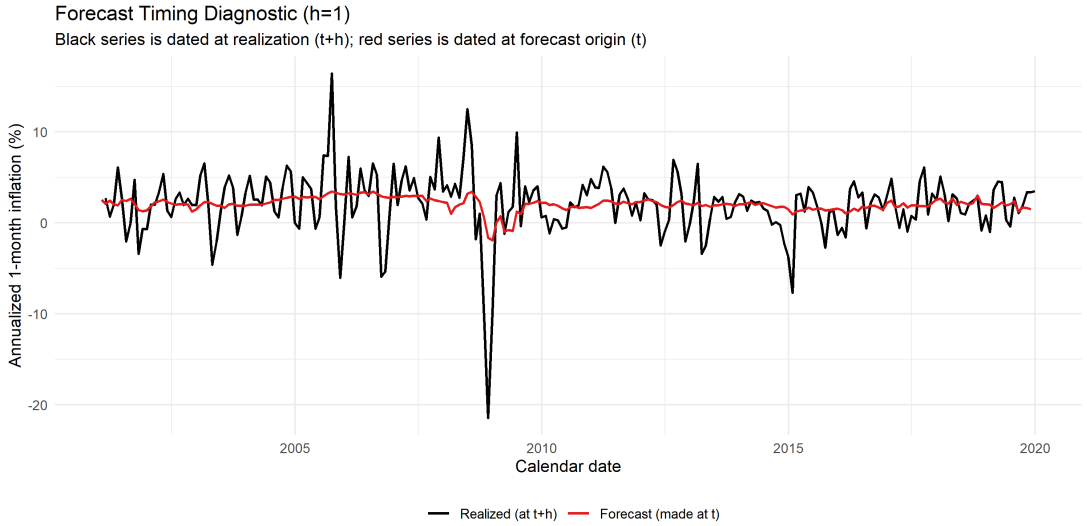

Figure 8: Forecast timing diagnostic ($h = 1$)

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

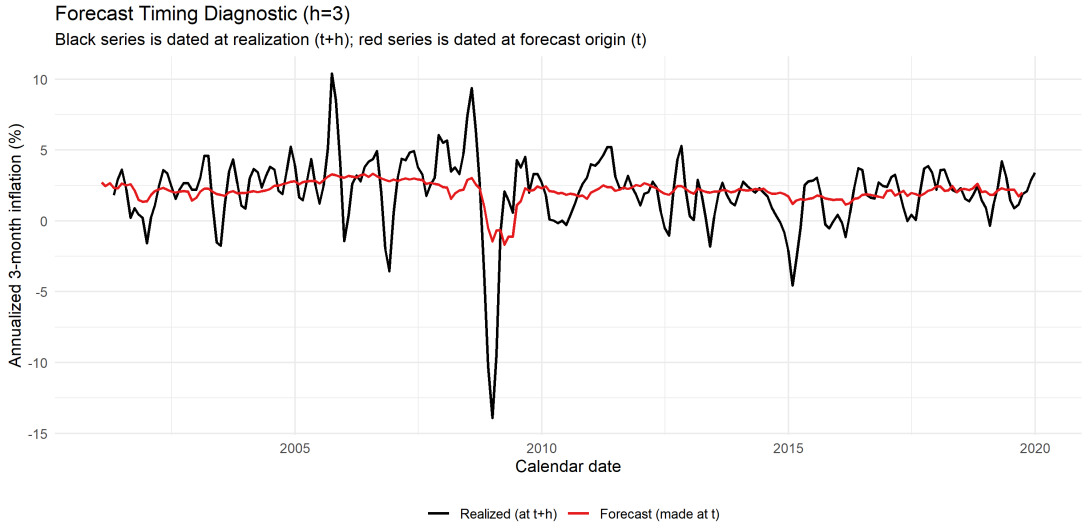

Figure 9: Forecast timing diagnostic ($h = 3$)

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

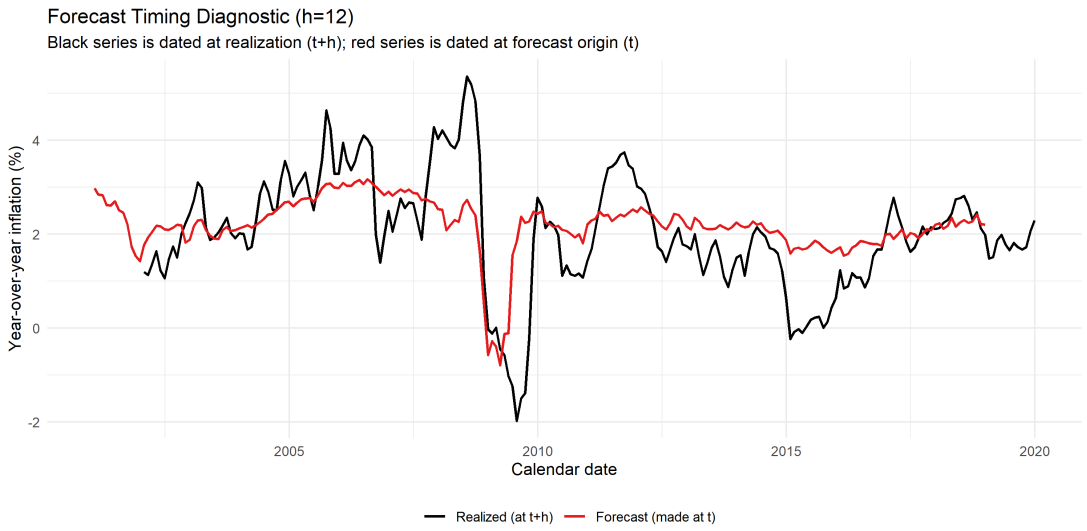

Figure 10: Forecast timing diagnostic ($h = 12$)

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

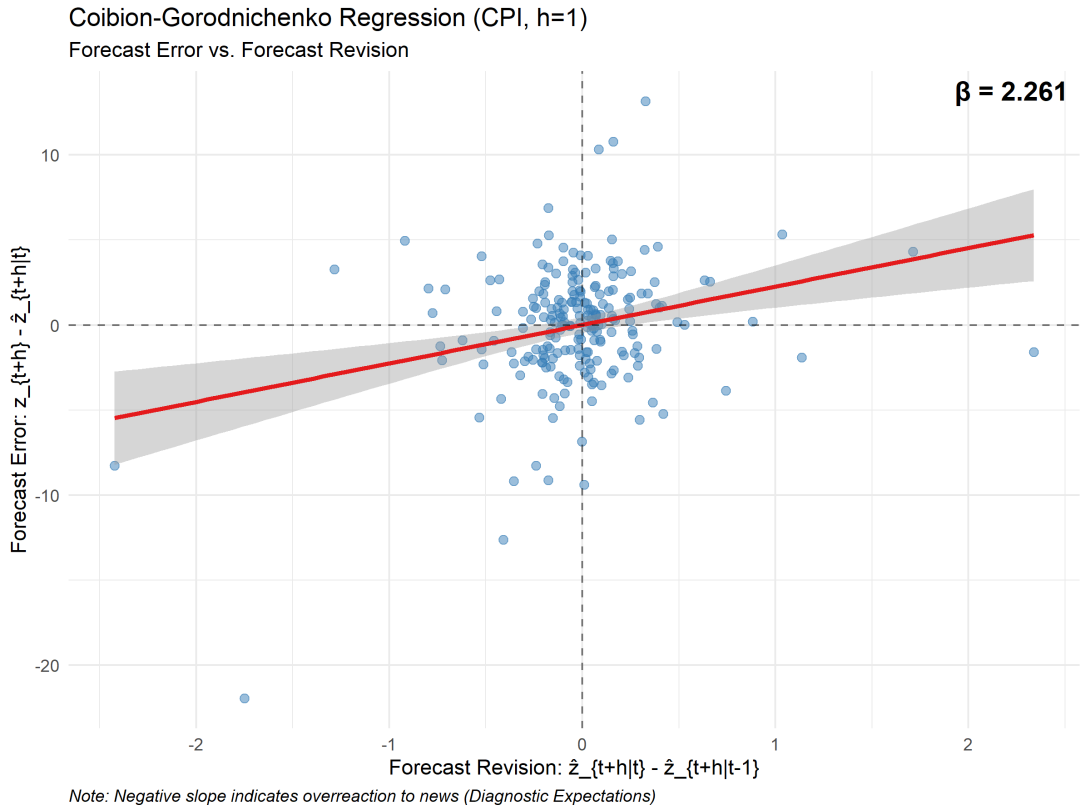

Figure 11: Revision diagnostic scatter: CPI ($h = 1$)

Notes: Scatter of forecast errors against forecast revisions for CPI inflation in the baseline design. The fitted line corresponds to the Coibion and Gorodnichenko (2015) regression.

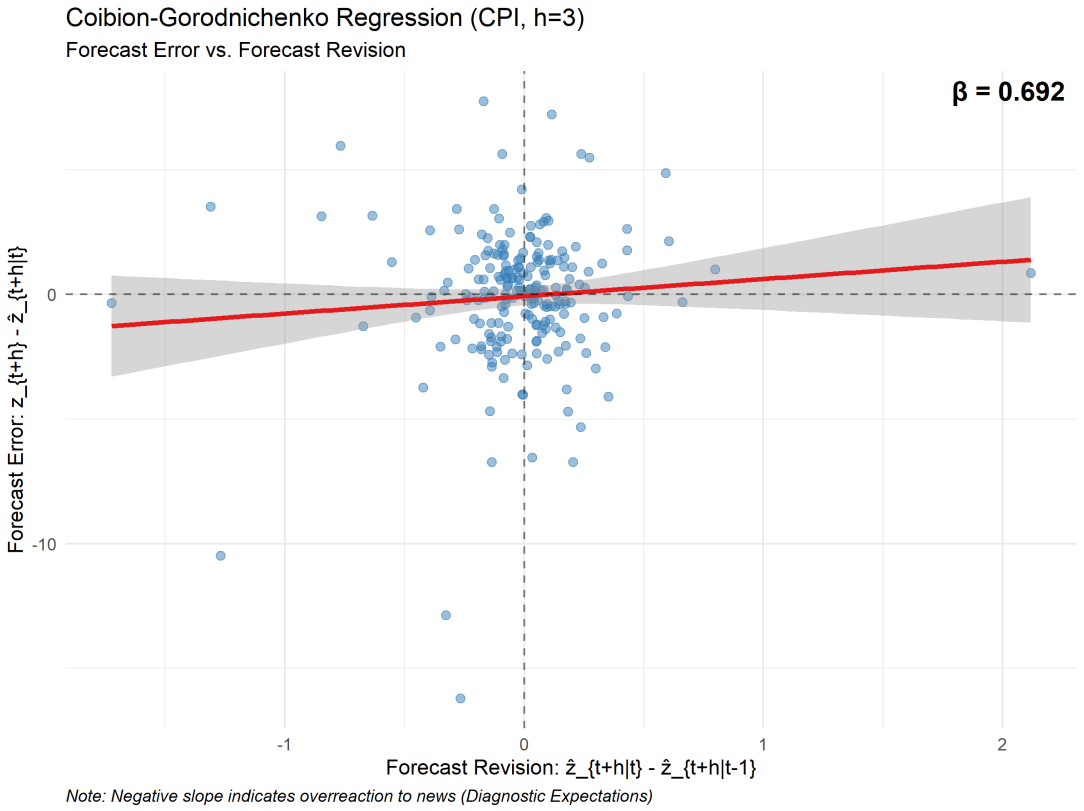

Figure 12: Revision diagnostic scatter: CPI ($h = 3$)

Notes: Scatter of forecast errors against forecast revisions for CPI inflation in the baseline design. The fitted line corresponds to the Coibion and Gorodnichenko (2015) regression.

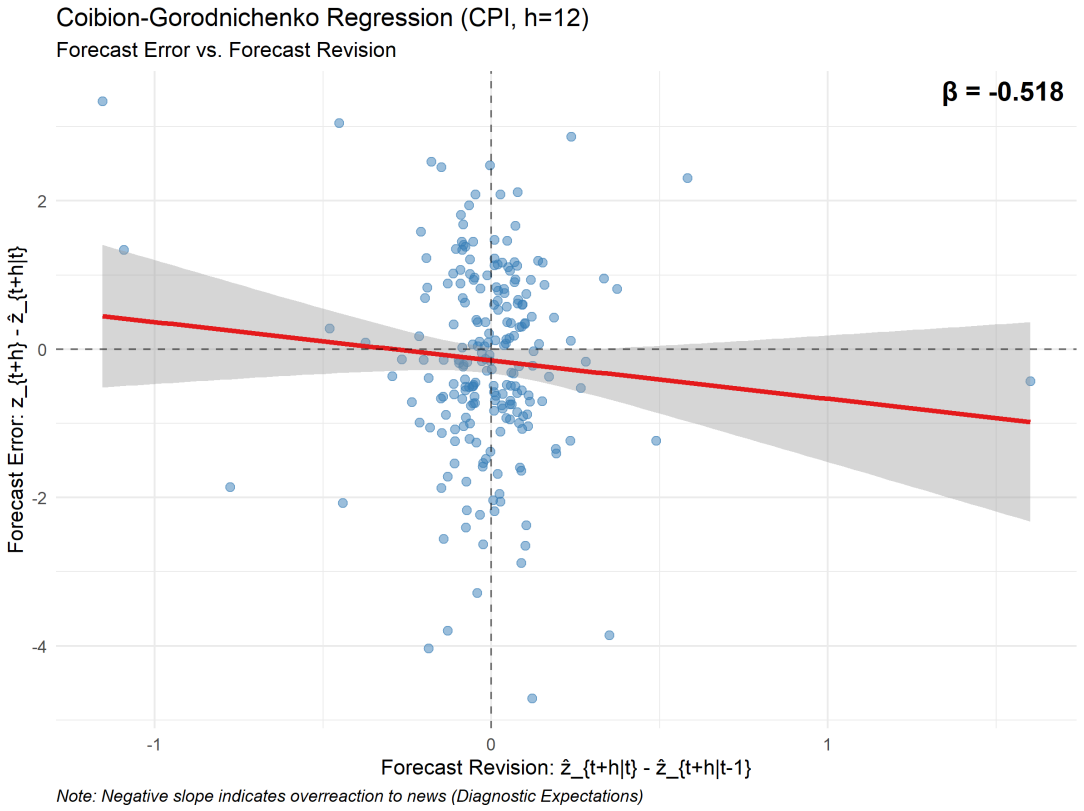

Figure 13: Revision diagnostic scatter: CPI ($h = 12$)

Notes: Scatter of forecast errors against forecast revisions for CPI inflation in the baseline design. The fitted line corresponds to the Coibion and Gorodnichenko (2015) 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`.