When Inflation Turns Volatile: Diverging Expectations of Households and Professionals*

Yi-Hsuan Chang Chien Nan-Kuang Chen Tsung-Hsien Li

This version: September 2026

Abstract

Are professional inflation forecasts more accurate than household expectations, and do the two groups form expectations differently across inflation environments? Using median one-year-ahead inflation expectations from the University of Michigan Survey of Consumers, the Survey of Professional Forecasters, and the Livingston Survey over 1981–2024, we compare household and professional forecasts across stable and volatile inflation regimes identified by a two-state Markov-switching model of realized consumer price inflation. We find no statistically significant evidence that either household or professional forecasts are systematically more accurate, either overall or within inflation regimes. The more distinctive result concerns expectation formation. In stable environments, households and professionals place similar cumulative weight on recently realized inflation. When inflation becomes volatile, their responses diverge: households place more weight on recent inflation, while professionals place less. These results are robust to the survey frequency, the regime definition, the standard-error estimator, and uncertainty in the estimated regime. Our results suggest that household and professional expectations provide different signals about inflation dynamics, supporting the joint monitoring of both measures by central banks when inflation is volatile.

Keywords: Inflation expectations; Survey forecasts; Markov-switching regimes; Forecast evaluation; Expectation formation; Inattention

JEL Classification: C22; C53; D84; E31; E37

*We thank participants at the 9th HenU/INFER Workshop on Applied Macroeconomics for helpful comments and Cheng-Yun Chen for excellent research assistance. **Chang Chien**: Research Assistant, Institute of Economics, Academia Sinica (email: ccyh@as.edu.tw). **Chen**: Professor, Department of Economics, National Taiwan University (email: nankuang@ntu.edu.tw). **Li** (corresponding author): Assistant Research Fellow, Institute of Economics, Academia Sinica (email: thli@econ.sinica.edu.tw). We gratefully acknowledge financial support from NSTC under Grant No. 112-2410-H-002-239.

1 Introduction

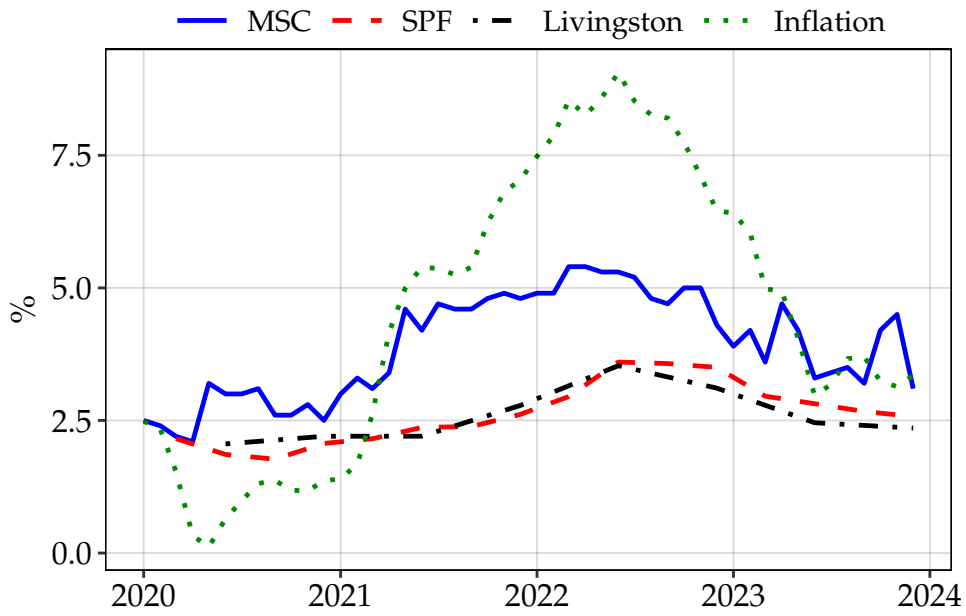
When prices in advanced economies rose abruptly in the early 2020s after three decades of relative stability, central banks faced a familiar question in an unfamiliar environment: what do people expect inflation to be? Expected inflation is embedded in the New Keynesian Phillips curve and monetary policy rules, and it also deeply affects wage bargaining, durable goods purchases, and portfolio decisions (Bernanke 2007; Lieb and Schuffels 2022; Burke and Ozdagli 2023). Expectations, however, are unobservable. Applied work must therefore choose an empirical proxy, often between surveys of households and surveys of professional forecasters. That choice typically rests on a familiar characterization of the two groups: household inflation expectations are upward-biased, dispersed, and historically less accurate, while professional forecasts are treated as a more informed measure of expected inflation (Ang, Bekaert, and Wei 2007; D'Acunto, Malmendier, and Weber 2023; Weber, D'Acunto, Gorodnichenko, and Coibion 2022; Henry, Mulloy, and Sarte 2023).

The 2021–2023 episode is awkward for that characterization. Figure 1 plots median one-year-ahead inflation expectations from the University of Michigan Survey of Consumers (MSC), the Survey of Professional Forecasters (SPF), and the Livingston Survey against realized CPI inflation. As inflation accelerated, household expectations rose earlier and considerably further than professional forecasts. The MSC median reached 5.3% in June 2022, when inflation peaked near 9.1%, whereas the professional medians reached only about 3.6% in the SPF and 3.5% in the Livingston Survey. Households, despite their reputation for inattention, responded strongly to the changing inflation environment, while professionals responded much less strongly.

An episode by itself does not establish that households became better forecasters than professionals. It raises a more fundamental question: does the relationship between household and professional inflation expectations change with the inflation environment? If the way agents acquire and process information changes with that environment, then differences between household and professional expectations may reflect not simply differences in forecast accuracy, but differences in how the two groups form their expectations.

Several theories imply precisely such state dependence. Sticky-information, rational-inattention, and noisy-information models suggest that agents update their beliefs based on the incentive to acquire and process information (Mankiw and Reis 2002; Sims 2003; Coibion and Gorodnichenko 2015). When inflation is stable and the cost of ignoring it is small, households may devote little attention to aggregate price developments. Conversely, when inflation begins moving sharply, the value of timely information rises, inducing households to track the economy more actively. Recent

Figure 1: Expected and Realized CPI Inflation, 2020–2023



Notes: The figure plots median one-year-ahead expected CPI inflation of households (MSC) and professional forecasters (SPF and Livingston Survey), together with realized year-over-year CPI inflation, January 2020–December 2023. Sources: University of Michigan, via the Federal Reserve Economic Data (FRED) database, and the Federal Reserve Bank of Philadelphia.

empirical evidence strongly supports this mechanism, demonstrating that household attention to inflation is highly sensitive to the prevailing inflation environment (Bracha and Tang 2025; Korenok, Munro, and Chen 2026; Weber, Candia, Afrouzi, Ropele, Lluberas, Frache, Meyer, Kumar, Gorodnichenko, Georgarakos, Coibion, Kenny, and Ponce 2025). Crucially, if household attention is state-dependent, differences between household and professional expectations may themselves depend on the inflation environment.

To evaluate this contingency, we examine how the relative accuracy and, more importantly, the formation of household and professional inflation expectations vary across inflation environments. We compare the median one-year-ahead inflation expectations from the MSC, the SPF, and the Livingston Survey over 1981–2024 in stable and volatile inflation environments identified by a two-state Markov-switching autoregression fitted to realized CPI inflation (Hamilton 1989; Evans and Wachtel 1993). Our conditioning variable is thus inflation volatility, classified endogenously rather than arbitrarily imposed as a threshold on the level. The innovation variance in the volatile state is roughly seven times that in the stable state. We date the regime at the forecast origin, so that it describes the inflation environment prevailing when expectations were formed, and examine forecast accuracy, unbiasedness, own-forecast efficiency, error persistence, the incorporation of public macroeconomic data, and the weight placed on recently realized inflation within each regime.

Our analyses yield three main findings. First, we find no statistically significant evidence that professional forecasts are systematically more accurate than household forecasts, or vice versa. Descriptively, professional forecasts exhibit lower root-mean-square error (RMSE) and mean absolute error (MAE) when inflation is stable, whereas household forecasts exhibit lower RMSE and MAE when inflation is volatile. Formal tests, however, do not establish either ranking. Across both professional comparison surveys, both loss functions, and both inflation regimes, none of the pairwise mean loss differences is statistically distinguishable from zero under Diebold and Mariano (1995) inference, with prewhitened Newey–West standard errors and moving-block bootstrap intervals. Conditional predictive ability tests (Giacomini and White 2006) similarly show no evidence that the inflation regime or past relative forecast performance predicts which survey will subsequently be more accurate. The fluctuation tests (Giacomini and Rossi 2010) reveal no subperiod in which one group achieves a statistically detectable advantage. Thus, although the point estimates suggest that relative forecast performance reverses across inflation environments, we cannot statistically establish either the conventional professional advantage or its apparent reversal during volatile periods. These tests may have limited power due to overlapping one-year-ahead errors and few volatile-regime common dates, so we describe the two sets of forecasts as statistically comparable rather than demonstrably equal.

Second, our rationality tests reveal state dependence in forecast inefficiency. In volatile inflation environments, professional forecast errors become strongly predictable from the professionals' own forecasts: the own-forecast coefficients are -0.831 for the SPF and -0.871 for Livingston, both significant at the 1% level. Household forecasts exhibit a similarly large point estimate in volatile periods. In contrast, the inefficiencies distinctive to households arise primarily in stable inflation environments: household forecast errors show positive autocorrelation and predictability from publicly available macroeconomic information, and we read both patterns as suggestive. Thus, departures from rational expectations do not map cleanly onto a distinction between "uninformed households" and "informed professionals". Instead, the nature of forecast inefficiency itself varies across inflation environments and across forecaster types.

Third, and most distinctively, households and professionals differ in how they form expectations, and these differences emerge precisely when inflation becomes volatile. When inflation is stable, the three surveys place remarkably similar cumulative weight on recently realized inflation. When inflation becomes volatile, however, their responses diverge. The cumulative loading on recent inflation increases for households, from 0.305 to 0.383, while it decreases for professional forecasters, from 0.284 to 0.242 for the SPF and from 0.302 to 0.193 for the Livingston Survey. The regime interaction is positive and statistically significant for households and negative for both professional

surveys, significantly so for Livingston. In other words, when inflation turns volatile, households become more responsive to recent inflation just as professional forecasts become less responsive to it.

The loss-difference tests, the professional inefficiency, and the cumulative loadings are robust to the survey frequency, the regime definition, the standard-error estimator, and the uncertainty in the estimated regime.

Taken together, these results reveal a distinction between forecast accuracy and expectation formation. Although we cannot statistically rank household and professional forecasts by accuracy, their expectation-formation processes respond differently when inflation turns volatile. We interpret this pattern as consistent with *state-dependent attention and extrapolation*. As inflation becomes more volatile, the returns to acquiring and processing inflation information increase, and household expectations become more responsive to recently observed inflation. Professional forecasters, who have stronger incentives to monitor macroeconomic conditions across inflation environments, instead place less weight on recent realized inflation during volatile periods, potentially relying more heavily on models, policy anchors, or other forward-looking information. We do not directly observe information acquisition, so we read state-dependent attention as an economic interpretation linking the observed changes to theories of costly information acquisition, not as an established causal mechanism.

Our paper makes three contributions. First, we shift attention from state dependence in relative forecast performance to state dependence in the underlying formation of household and professional expectations. Prior research shows that forecast performance can vary with economic conditions, and recent evidence shows that household attention to inflation is itself state-dependent. We connect these literatures by examining households and professional forecasters within a common empirical framework. In stable inflation environments, the two groups exhibit nearly identical backward-looking loadings. When inflation becomes volatile, however, their responses move in opposite directions: households place more weight on recent inflation while professionals place less. Thus, changes in the inflation environment affect not merely which forecast appears more accurate, but the relationship between the expectation-formation processes represented by household and professional surveys.

Second, we provide a comprehensive reassessment of the presumed professional advantage in inflation forecasting. We complement conventional RMSE and MAE comparisons with unconditional loss-difference tests, conditional predictive-ability tests, and rolling-window fluctuation tests for unstable forecasting environments. Across these approaches, we cannot establish a professional accuracy advantage in either stable or volatile inflation regimes. This finding does not imply equal accuracy in pop-

ulation; rather, it shows that the presumed professional advantage is not statistically supported in our sample. The accuracy results therefore motivate our main finding: if neither group can be reliably ranked by forecast performance, the economically relevant distinction lies in how their expectations are formed.

Third, we show that the relevant state dependence is associated with inflation turbulence rather than simply with high inflation. Rather than defining environments using *ex ante* high- and low-inflation thresholds, we identify stable and volatile regimes endogenously from the inflation process. The resulting classification captures periods of large inflation movements in either direction, including rapid disinflation, rather than simply periods of high inflation.

These findings have implications for how central banks assess inflation expectations and potential movements in inflation. When inflation becomes volatile, professional forecasts become less responsive to recently realized inflation while household expectations become more responsive, even though neither group can be shown to forecast more accurately. Thus, the two measures may provide different signals about inflation dynamics rather than alternative estimates of the same latent expectation. For central banks, relying primarily on professional forecasts may understate the extent to which recent inflation is being incorporated into household expectations. Monitoring both measures, and particularly their divergence, may therefore provide a more complete assessment of how inflationary developments are propagating through expectations and of the potential persistence of inflation. More broadly, the choice of expectation measure in Phillips curves, monetary-policy rules, and other applications should be treated as a modeling assumption rather than based on a presumed ranking of forecast accuracy.

The empirical evidence of this paper relates to several strands of literature. A long literature evaluates the accuracy and rationality of survey-based inflation expectations. Mehra (2002) finds that MSC expectations exhibit greater evidence of bias over 1961–2000, whereas Livingston Survey forecasts are broadly unbiased and efficient. Ang et al. (2007) show that survey forecasts generally outperform time-series and asset-market alternatives, with professional surveys modestly outperforming household expectations. Related research documents the upward bias, dispersion, and volatility of household inflation expectations (D'Acunto et al. 2023; Weber et al. 2022) and their historically weaker average forecast performance (Henry et al. 2023). Our evidence challenges this conventional ranking by showing that household and professional forecasts cannot be statistically ranked once the performance differences are formally tested, either unconditionally or conditional on the inflation environment.

Our analysis also relates to research showing that relative forecast performance can

vary across economic environments. Diebold and Mariano (1995) provide the framework for testing unconditional differences in predictive accuracy, while Giacomini and White (2006) develop tests of conditional predictive ability and Giacomini and Rossi (2010) show how forecast rankings can change over time. Recent applications provide further evidence that relative forecast performance depends on economic conditions. Candelon and Roccazzella (2025), for example, document changes in relative inflation-forecast performance as macroeconomic uncertainty and financial conditions shifted. Related evidence also cautions against treating professional forecasts as a uniformly superior benchmark. El-Shagi (2011) shows that survey-based inflation expectations can perform comparably to sophisticated econometric forecasts, while Benchimol, El-Shagi, and Saadon (2022) document substantial heterogeneity in forecasting behavior and performance even among professional forecasters. Goodspeed (2025), most closely related to our setting, finds that the professional advantage over households is concentrated in low-inflation periods. We examine a different dimension of the inflation environment: volatility in the inflation process rather than the level of inflation. The distinction matters because periods in which inflation is difficult to track need not coincide with periods in which inflation is high. Because we identify stable and volatile regimes endogenously through a Markov-switching model, our classification captures large inflation movements in either direction, including rapid disinflation, that a high-versus-low inflation split can miss. We then ask whether the underlying expectation-formation processes of households and professionals themselves respond differently to the inflation environment.

Finally, our results speak to theories of costly information acquisition and state-dependent attention. Sticky-information, rational-inattention, and noisy-information models imply that agents' information-processing decisions respond to the benefits of acquiring and incorporating new information (Mankiw and Reis 2002; Sims 2003; Coibion and Gorodnichenko 2015). Mankiw, Reis, and Wolfers (2003) provide the classic empirical benchmark: median expectations from the same three surveys we use are consistent with neither full rationality nor pure adaptive expectations. Recent evidence provides direct support for such state dependence among households. Bracha and Tang (2025) show that household attention increases with inflation; Korenok et al. (2026) document inflation-attention thresholds across countries; and Weber et al. (2025) show that exogenously provided inflation information has smaller effects when inflation is already elevated and respondents arrive better informed. Households also place disproportionate weight on salient and personally experienced prices (Malmendier and Nagel 2016; D'Acunto, Malmendier, Ospina, and Weber 2021), while individual professional forecasters can overreact to news (Bordalo, Gennaioli, Ma, and Shleifer 2020). We extend this literature by bringing household and professional expectations

into the same state-dependent framework: we take the classic rationality and adaptiveness tests and let every coefficient differ between stable and volatile regimes. The professional response to volatile inflation is not simply a muted version of the household response. Instead, the two groups move in opposite directions: professionals place less weight on recent realized inflation as households place more. This divergence shows that state-dependent household attention alone cannot characterize how different expectation measures respond to inflation turbulence and points to systematic differences in how households and professionals process and weight information when inflation becomes difficult to forecast.

More broadly, measured inflation expectations enter both economic decisions and policy analysis: expected inflation affects household spending and portfolio choices and plays a central role in monetary-policy transmission (Bernanke 2007; Lieb and Schuffels 2022; Burke and Ozdagli 2023). Central-bank communication can also directly alter household expectations (Coibion, Gorodnichenko, and Weber 2022). Reis (2021) emphasizes that assessing whether inflation expectations have become unanchored requires comparing multiple expectation measures rather than relying on a single series. Our results provide an additional reason for doing so: household and professional expectations can respond differently to the same change in the inflation environment even when their realized forecast accuracy remains statistically comparable. Treating either measure as "the" inflation expectation can therefore obscure economically meaningful differences in how expectations are formed precisely when inflation is most volatile and expectations are most consequential.

The remainder of the paper proceeds as follows. Section 2 describes the survey data, forecast measures, and empirical methodology. Section 3 presents the inflation-regime classification. Section 4 evaluates relative forecast accuracy across regimes. Section 5 examines forecast rationality and expectation formation. Section 6 reports robustness analyses, including alternative regime definitions and inference accounting for regime-estimation uncertainty. Section 7 concludes. Appendices A–E collect supplementary results.

2 Data, Forecast Measures, and Methodology

We use the median-based one-year-ahead expected rate of consumer price inflation from three long-running U.S. surveys: the monthly MSC for households, and the quarterly SPF and semiannual Livingston Survey for professionals. The sample starts in 1981, when the SPF begins asking about CPI inflation, and the evaluation window ends with 2024, the last calendar year of forecasts with realized one-year-ahead out-

comes. Alongside the survey medians, the analysis uses realized CPI inflation as the target and the three-month Treasury bill rate and the civilian unemployment rate for testing the use of macroeconomic information. Appendix A collects the data details.

Section 2.1 describes the three surveys and the realized-inflation target. Sections 2.2–2.3 collect the paper's core specifications.

2.1 Survey measures of inflation expectations

The MSC has run since 1946, monthly since 1978, on a rotating panel of roughly 500 respondents. The survey asks whether prices in general will rise, fall, or stay the same over the next twelve months and "by about what percent". We use the published median of the point responses, the most widely used series and the only one available monthly over the full sample.[1]

The SPF, run by the American Statistical Association and the National Bureau of Economic Research (NBER) from 1968 and by the Federal Reserve Bank of Philadelphia since 1990, surveys roughly 30–40 professional forecasters each quarter. The SPF's one-year-ahead CPI question begins in 1981:Q3. The Livingston Survey began in 1946, and the Philadelphia Fed has likewise run it since 1990. The survey's roughly 20–30 respondents are economists spanning academia, banking, government, and industry.

Realized inflation π_t is the year-over-year growth rate of the U.S. CPI for All Urban Consumers, not seasonally adjusted. Each survey yields a median-based expected inflation rate over the next twelve months. The target of a forecast formed at t is therefore π_{t+12} in monthly time and the forecast error is

$$e_{t+h} \;=\; \pi_{t+h} - \mathbb{E}_t \pi_{t+h}, \qquad h = 12 \text{ months}, \tag{1}$$

so a negative error means the survey over-predicted inflation and a positive error means it under-predicted inflation. We call t the *forecast origin* and $t + h$ the *realization date*.

The paper uses two sample windows.

1. The evaluation window runs from the first common survey date through the last calendar year of forecasts with realized one-year-ahead outcomes: 1981:M9–2024:M12 for the MSC (520 months), 1981:Q3–2024:Q4 for the SPF (174 quarters), and 1981:H2–2024:H2 for the Livingston Survey (87 half-years).[2]

[1] The New York Fed's Survey of Consumer Expectations elicits richer objects but begins only in June 2013, too short for regime-conditional analysis.

[2] The October 2025 CPI was never published owing to the 2025 U.S. federal government shutdown, so MSC error samples lose the October 2024 forecast, leaving 519 rather than 520 months.

2. We estimate the Markov-switching model over 1981:Q2–2025:Q4. The first observation conditions the autoregression, so regime probabilities run from 1981:Q3; we call 1981:Q3–2025:Q4 the classification window. The estimation sample extends one year past the evaluation window to include every realized value needed for the forecast-error evaluation.

2.2 A two-state Markov-switching model of inflation

We recover inflation regimes from the inflation process itself rather than imposing a threshold, estimating a two-state Markov-switching autoregression of order one for quarterly year-over-year CPI inflation:

$$\pi_t = \alpha_{s_t} + \rho_{s_t}\,\pi_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_{s_t}\right), \qquad s_t \in \{1, 2\}, \tag{2}$$

where the latent state s_t follows a first-order Markov chain with transition probabilities $p_{ij} = \Pr(s_t = j \mid s_{t-1} = i)$ and implied steady-state mean $\mu_s = \alpha_s/(1-\rho_s)$. State-dependent intercepts and variances let the states differ in the level and variability of inflation. We impose two states for parsimony and interpretability.[3] We estimate the model by maximum likelihood with the expectation–maximization algorithm on the 1981:Q2–2025:Q4 sample above.

We ask how inflation expectations differ across households and professionals, and how each group forms them, conditional on the inflation environment prevailing when the forecast is made. We characterize that environment *ex post* because the full sample gives the best available description of the conditions each forecaster faced. To this end, we define the regime dummy Regime_t as an indicator that the smoothed probability of state 2 exceeds one-half:

$$\text{Regime}_t \;=\; \mathbb{1}\{\Pr(s_t = 2 \mid \mathcal{I}_T) > 0.5\}, \tag{3}$$

where $\mathcal{I}_T$ is the full estimation sample. We date the regime dummy at the forecast origin t, the date at which the expectation is formed. Throughout, state 1 is the stable regime and state 2 the volatile regime.[4]

[3] For example, Evans and Wachtel (1993) model U.S. inflation as switching between two Markov regimes when decomposing inflation uncertainty, and Goodspeed (2025) evaluates household and professional forecast performance across two inflation states.

[4] The ordering is a normalization, because equation (2) identifies the states only up to relabeling. The names are not: state 2 has the higher variance and the higher mean in Table 1.

2.3 Evaluation specifications

We evaluate each survey over the full sample and within each regime on (1) accuracy, (2) four dimensions of rationality (unbiasedness, own-forecast efficiency, error persistence, and use of macroeconomic information), and (3) adaptiveness. Every specification interacts the object of interest with Regime_t. Baseline coefficients give stable-regime effects and baseline-plus-interaction sums give volatile-regime effects.

Accuracy. We measure accuracy by the RMSE and MAE of each survey's errors and, formally, by pairwise loss differences:

$$d_t = \mathcal{L}\left(e_{t+h}^{MSC}\right) - \mathcal{L}\left(e_{t+h}^{Prof}\right), \qquad \mathcal{L}(e) \in \{e^2,\ |e|\}, \tag{4}$$

where e_{t+h}^{S} is survey S's error on the one-year-ahead forecast made at t, so the loss difference is indexed by the forecast origin. We compute the loss differences on common dates, and a negative d_t indicates lower household loss relative to professionals.

In the spirit of Diebold and Mariano (1995), we test whether the mean loss difference is zero, overall and within each regime, with the regression:

$$d_t \ = \ \mu_S\,(1 - \text{Regime}_t) + \mu_V\,\text{Regime}_t + \varepsilon_t, \tag{5}$$

with heteroskedasticity- and autocorrelation-consistent (HAC) inference on μ_S and μ_V and moving-block bootstrap intervals as a complement. Imposing $\mu_S = \mu_V$ (a single constant in place of the two regime indicators) gives the full-sample test of $\mathbb{E}[d_t] = 0$.

Rationality. The four rationality tests share one benchmark: under full-information rational expectations, forecast errors are unbiased and orthogonal to the period-t information set. The specifications follow Mankiw et al. (2003), and we extend them with the regime interactions.

Unbiasedness. We regress the error on a constant and the regime dummy:

$$e_{t+h} = \alpha_1 + \alpha_2 \cdot \text{Regime}_t + \varepsilon_{t+h}, \tag{6}$$

so α_1 is the mean error in the stable regime and $\alpha_1 + \alpha_2$ the mean in the volatile regime. Under unbiasedness, both means are zero, and a negative mean indicates systematic over-prediction.

Own-forecast efficiency. We add the forecast itself:

$$e_{t+h} = \alpha_1 + \beta_1 \mathbb{E}_t[\pi_{t+h}] + (\alpha_2 + \beta_2 \mathbb{E}_t[\pi_{t+h}]) \cdot \text{Regime}_t + \varepsilon_{t+h}. \tag{7}$$

A nonzero β_1, or volatile-regime sum $\beta_1 + \beta_2$, signals unexploited information.

Error persistence. We replace the forecast with the survey's most recent own error, $e_t = \pi_t - \mathbb{E}_{t-12}[\pi_t]$:

$$e_{t+h} = \alpha_1 + \beta_1 e_t + (\alpha_2 + \beta_2 e_t) \cdot \text{Regime}_t + \varepsilon_{t+h}. \tag{8}$$

A nonzero β_1 or sum $\beta_1 + \beta_2$ means past errors predict the current one.

Use of macroeconomic information. We extend equation (7) with lagged inflation π_{t-1}, the lagged three-month Treasury bill rate i_{t-1}, and the lagged civilian unemployment rate $\mathcal{U}_{t-1}$. The bill rate and the unemployment rate are available monthly in real time and enter lagged one month, for all three surveys, so they sit unambiguously in the information set at the forecast origin:

$$\begin{aligned} e_{t+h} = {} & \alpha_1 + \beta_1 \mathbb{E}_t[\pi_{t+h}] + \gamma_1 \pi_{t-1} + \kappa_1 i_{t-1} + \delta_1 \mathcal{U}_{t-1} \\ & + (\alpha_2 + \beta_2 \mathbb{E}_t[\pi_{t+h}] + \gamma_2 \pi_{t-1} + \kappa_2 i_{t-1} + \delta_2 \mathcal{U}_{t-1}) \cdot \text{Regime}_t + \varepsilon_{t+h}. \end{aligned} \tag{9}$$

We focus on the joint use of macroeconomic information: a rational forecast embeds these public series, so γ_1, κ_1, and δ_1 should be jointly zero. The volatile-regime sums $\gamma_1 + \gamma_2$, $\kappa_1 + \kappa_2$, and $\delta_1 + \delta_2$ should likewise be jointly zero.

Expectation adaptiveness. We regress the survey forecast on the twelve monthly lags of realized inflation preceding the forecast origin, with regime-varying coefficients, controlling for current and lagged unemployment and interest rates:

$$\mathbb{E}_t[\pi_{t+h}] = \alpha + \left(\beta_1(L) + \beta_2(L) \cdot \text{Regime}_t\right) \pi_t + \lambda_1 \mathcal{U}_t + \lambda_2 \mathcal{U}_{t-1} + \theta_1 i_t + \theta_2 i_{t-1} + \varepsilon_t, \tag{10}$$

where L is the lag operator and $\beta_k(L) = \sum_{j=1}^{12} \beta_{kj} L^j$ for $k = 1, 2$. Our interest is in the sum of the lag coefficients, not the individual terms. We write $\beta_k(1) = \sum_{j=1}^{12} \beta_{kj}$ for the *cumulative loading*: the total change in the forecast, once all twelve lags have acted, per percentage point of realized inflation. The cumulative loading is $\beta_1(1)$ in the stable regime and $\beta_1(1) + \beta_2(1)$ in the volatile regime.

Inference. Forecast errors here are serially correlated by construction, because successive one-year forecast windows overlap. All standard errors in the survey evaluations are therefore *prewhitened* Newey–West estimators (Newey and West 1987; Andrews and Monahan 1992) with a bandwidth of one year at each survey's frequency (twelve, four, and two lags, so every bandwidth covers the overlap). Every evaluation quantity carries a moving-block bootstrap complement with one-year blocks, with the

Table 1: Estimated Two-State Markov-Switching Model of CPI Inflation

	$\widehat{\alpha}_s$	$\widehat{\rho}_s$	$\widehat{\sigma}_s^2$	$\widehat{\mu}_s$	$\widehat{p}_{ss}$
Regime 1: Stable	0.282	0.873	0.211	2.212	0.950
Regime 2: Volatile	0.713	0.807	1.573	3.693	0.899

Notes: Maximum-likelihood estimates of the MS-AR(1) model in equation (2), fitted to quarterly year-over-year CPI inflation over 1981:Q2–2025:Q4 (the first observation serves as the conditioning value, so regime probabilities begin in 1981:Q3; $N = 178$). $\widehat{\mu}_s = \widehat{\alpha}_s/(1-\widehat{\rho}_s)$ is the implied steady-state mean of inflation in regime s, and $\widehat{p}_{ss} = \Pr(s_{t+1} = s \mid s_t = s)$ the estimated probability of remaining in regime s next quarter; the off-diagonal transition probabilities are $1 - \widehat{p}_{ss}$ (0.050 and 0.101), so both regimes are highly persistent. Standard errors of the intercept and autoregressive parameters: 0.087 and 0.031 (Regime 1), 0.173 and 0.038 (Regime 2).

resampling mechanics detailed in Appendix E.3.

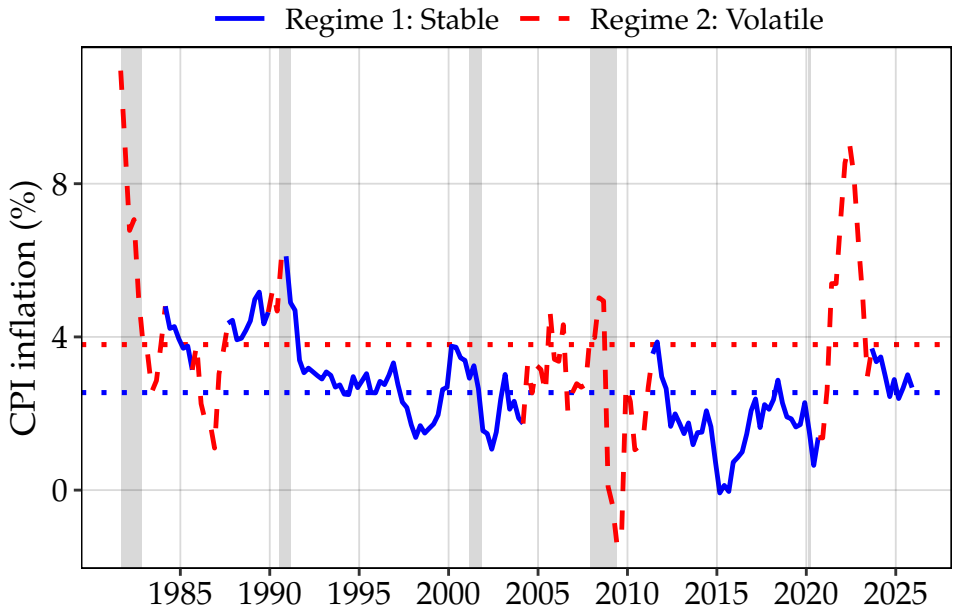
Robustness checks. In Section 6, we recompute the household loadings on the professional surveys' observation dates, vary the cutoff behind the regime indicator, replace smoothed with filtered probabilities, compare the prewhitened standard errors with their non-prewhitened counterparts, and use a parametric bootstrap to measure how much the uncertainty from estimating the regime indicator carries into the results.

3 Inflation Regime Identification

Table 1 reports the estimates of the Markov-switching model of Section 2.2. Regime 1 combines a steady-state mean of $\widehat{\mu}_1 = 2.212$ with a small innovation variance of 0.211. Regime 2 combines a mean of $\widehat{\mu}_2 = 3.693$ with an innovation variance of 1.573, roughly seven times larger. Both states are highly persistent, with continuation probabilities $\widehat{p}_{11} = 0.950$ and $\widehat{p}_{22} = 0.899$ and expected durations of about five years and about two and a half years. We name the states *stable* and *volatile*, not low- and high-inflation, because the variance gap is what separates them: the volatile state captures large inflation movements, rising or falling, rather than only "high inflation".

Figure 2 displays realized inflation colored by the classification of equation (3): five volatile episodes covering 64 of the 178 quarters, with nine transitions. The five episodes are the early-1980s disinflation (1981:Q3–1984:Q1), 1985:Q4–1987:Q3, 1990:Q1–1990:Q4, 2004:Q2–2011:Q2 (the pre-crisis commodity run-up through the financial crisis and the European sovereign-debt period), and the post-pandemic surge (2020:Q4–2023:Q3). The classification is mainly about the second moment of inflation, not just the first. Thus, Regime_t measures the turbulence of the inflation environment: how hard inflation is to track, not only how high it stands. The figure also shows what a level classification would miss. A threshold near 3% would label many volatile quarters as

Figure 2: Realized Inflation Classified by Inflation Regime



Notes: Blue (solid) denotes the stable regime (Regime 1) and red (dashed) the volatile regime (Regime 2); a period is classified as volatile when the smoothed probability of the volatile state exceeds 0.5. The two horizontal dotted lines mark the mean of realized inflation within each regime, in the matching color. Shaded bars denote NBER recessions.

low-inflation periods, such as the years after the financial crisis.[5]

Appendix B tests the premise that inflation variance is an informative basis for classifying inflation environments, with three procedures independent of the Markov-switching model. The parameter-constancy test of Nyblom (1989) and Hansen (1992) locates the instability in the innovation variance, not the conditional mean, most sharply at monthly frequency. The cumulative-sum-of-squares test of Inclán and Tiao (1994), corrected for conditional heteroskedasticity as in Sansó, Aragó, and Carrion-i-Silvestre (2004), rejects a constant variance at the 1% level on monthly data, so the variance regime is not merely generalized autoregressive conditional heteroskedasticity (GARCH). An iterative nonparametric search confirms the timing as well, placing its two change points three quarters inside the transitions that open and close the model's longest volatile episode. Taken together, the three procedures support reading the inflation environment off the variance. Two further checks in Appendix B corroborate the reading: the regime label predicts the size of inflation innovations, including through the post-2021 surge, and re-estimating the model on core inflation, which excludes food and energy, preserves the two episodes that dominate the evaluation window, so energy accounts for the middle of the sample rather than its ends.

[5]Of the 64 volatile quarters, 23 have inflation below the evaluation-window median of 2.8%, and a 4% level rule recovers only 24 of the 64. The standard deviation of the quarterly change in inflation, by contrast, is 1.38 percentage points in the volatile regime against 0.47 in the stable one.

4 Is There a Professional Accuracy Advantage?

This section comprehensively evaluates whether professional forecasters possess a systematic accuracy advantage over households. We find no statistically significant evidence of a forecast accuracy advantage for either group. Both the professionals' descriptive edge in the stable regime and the households' edge in the volatile regime dissolve under formal testing. Section 4.1 documents the pattern in the forecast errors, and Section 4.2 presents formal pairwise tests of equal predictive accuracy. Section 4.3 extends the analysis with a conditional predictive ability test and a window-by-window fluctuation test. Section 4.4 explains why neither edge survives testing and what that implies for the choice among surveys. In Appendix C, we specify the two tests of Section 4.3 and probe the no-advantage answer for significance levels and direction.

4.1 Descriptive accuracy across regimes

Table 2 reports the RMSE and MAE of one-year-ahead forecast errors for each survey. Over the full sample, the three surveys are close to indistinguishable: the MSC attains a slightly lower RMSE and the professional surveys a slightly lower MAE, so no survey dominates on both criteria.

The regime decomposition produces the pattern that motivates the paper. In the stable regime, the professional surveys are more accurate on both criteria: RMSEs of 1.04 for the SPF and 1.01 for Livingston against 1.21 for households. In the volatile regime, the point estimates reverse on both criteria: household RMSE of 2.08 against 2.26 for both professional surveys. The professional stable-regime edge is 0.17–0.20 RMSE points, roughly 15% of the household RMSE. The household volatile-regime edge is 0.17–0.18 RMSE points, roughly 8% of the professional RMSE. Whether either edge is distinguishable from zero, RMSE tables cannot say.

4.2 Tests of equal predictive accuracy

We test whether the mean of the pairwise loss difference d_t of equation (4) is zero, overall and within each regime, following equation (5). Table 3 covers two loss functions, squared and absolute, two comparisons, MSC–SPF and MSC–Livingston, and three samples, the full sample and the stable and volatile regimes, twelve cells in all. Each cell reports the mean loss difference, its prewhitened Newey–West standard error in parentheses, and a moving-block bootstrap 95% interval in brackets. The rule is conservative and symmetric: one survey outperforms the other only if an interval lies entirely on one side of zero. Otherwise the two are *comparable* in that regime.

Table 2: Forecast Errors

		MSC	SPF	Livingston
Full Period	RMSE	1.586	1.597	1.589
	MAE	1.168	1.085	1.093
Regime 1	RMSE	1.208	1.037	1.010
	MAE	0.932	0.784	0.780
Regime 2	RMSE	2.085	2.256	2.261
	MAE	1.576	1.603	1.632

Notes: RMSE and MAE of the one-year-ahead forecast errors of equation (1), in percentage points. Samples: MSC 519 monthly (1981:M9–2024:M12), SPF 174 quarterly, Livingston 87 semiannual; the October 2024 MSC observation is excluded because its realization (the October 2025 CPI) was never published owing to the 2025 U.S. federal government shutdown. Regime 1 (stable) and Regime 2 (volatile) date the regime at the forecast origin (Regime_t).

Table 3: Pairwise Loss-Difference Tests by Inflation Regime

Comparison	Full sample	Stable regime	Volatile regime
MSC − SPF, squared loss	0.038 (0.877) [−0.962, 1.137]	0.405 (0.410) [−0.173, 0.993]	−0.592 (2.185) [−3.120, 2.233]
MSC − SPF, absolute loss	0.088 (0.130) [−0.099, 0.304]	0.163 (0.140) [−0.041, 0.370]	−0.040 (0.233) [−0.406, 0.381]
MSC − Livingston, squared loss	0.009 (0.628) [−1.035, 1.168]	0.357 (0.321) [−0.131, 0.885]	−0.588 (1.538) [−3.327, 2.397]
MSC − Livingston, absolute loss	0.083 (0.117) [−0.105, 0.299]	0.161 (0.141) [−0.022, 0.356]	−0.050 (0.201) [−0.444, 0.401]

Notes: Entries are mean loss differences $\bar{d}$ as defined in equation (4); negative entries indicate lower loss for the MSC. Prewhitened Newey–West standard errors in parentheses (bandwidth of one year: four lags for the quarterly MSC–SPF comparison, two for the semiannual MSC–Livingston comparison) and moving-block bootstrap 95% percentile intervals in brackets (9,999 replications; block lengths of four and two periods, resampling (d_t, Regime_t) jointly). Regimes are dated at the forecast origin (Regime_t). Common dates: 174 quarterly observations (110 stable, 64 volatile) for MSC–SPF and 87 semiannual observations (55 stable, 32 volatile) for MSC–Livingston. No entry is significantly different from zero at the 10% level under Newey–West inference (smallest $p = 0.247$), and no 95% bootstrap interval excludes zero. Table C.1 in Appendix C reports the 90% intervals and the one-sided bootstrap tail areas.

For both professional comparisons and both loss functions, the point estimates track the descriptive pattern: positive in the stable regime, negative in the volatile. None of the twelve mean differences is statistically distinguishable from zero. The largest point estimate in absolute value, the MSC–SPF squared-loss difference of -0.592 in the volatile regime, carries a standard error of 2.185 and a bootstrap interval of $[-3.120, 2.233]$.

4.3 Conditional and window-by-window accuracy tests

The tests of Section 4.2 compare regime means, and a mean can conceal two failures of equal accuracy. First, relative performance may be predictable from information available at the origin, in which case a zero mean hides a usable switching rule. Second, one survey may dominate over a subperiod too short to move the mean. We take the two in turn, with the conditional predictive ability test of Giacomini and White (2006) and the fluctuation test of Giacomini and Rossi (2010). Appendix C specifies both, in equations (C.1) and (C.2).

Does anything known at the origin predict relative accuracy? The conditional test regresses the loss difference on instruments dated at the forecast origin and asks whether they jointly predict it. Those instruments must be measurable with respect to the origin's information set, and that requirement bites on our own regime indicator. The baseline of equation (3) is built from smoothed probabilities, which condition on the full sample, so it is not measurable at the origin. We therefore run every test twice, once with the baseline indicator and once with the filtered indicator $\text{Regime}_{f,t}$, which conditions only on inflation realized through t. Across the sixteen tests of Table 4, no Wald statistic rejects. The smallest Newey–West p-value is 0.140 and the smallest bootstrap p-value 0.255.

Making the instrument measurable moves the tests away from rejection, not toward it. With the regime alone as the instrument, replacing the baseline indicator with the filtered one raises the p-value from 0.580 to 0.927 for the MSC–SPF squared-loss comparison and from 0.483 to 0.915 for MSC–Livingston. The reason is that filtering relabels about one date in eight, and the relabeled dates track the loss difference no better than noise. The no-advantage answer therefore does not rest on the look-ahead in the classification, since even the indicator that draws on the full sample predicts nothing. The filtered indicator is still not fully real time, because its probabilities condition only on inflation realized through t while the model parameters behind them come from the full sample. A fully recursive construction, re-estimating those parameters at every origin, would add estimation noise to the relabeling noise, and the comparison above shows that such noise pushes the p-values up, not down. Dating the filtered indicator one quarter before the origin, which removes even the origin quarter's information, likewise changes no conclusion: the smallest Newey–West p-value across the eight tests is then 0.210 and the smallest bootstrap p-value 0.316.

Is the no-advantage answer an average over offsetting episodes? The fluctuation test computes the loss difference over rolling windows and compares the largest standardized value along the path with a critical value for the path as a whole. The MSC–SPF squared-loss path rises as far as 1.78 while the windows fill with the stable 2010s, fa-

Table 4: Giacomini–White Conditional Predictive Ability Tests

	MSC vs. SPF			MSC vs. Livingston		
Instruments	W	p_{NW}	p_{MBB}	W	p_{NW}	p_{MBB}
Panel A: squared loss						
$\{1,\ \text{Regime}_t\}$	1.091	0.580	0.580	1.456	0.483	0.559
$\{1,\ \text{Regime}_t,\ d_{t-s}\}$	1.055	0.788	0.809	2.460	0.483	0.576
$\{1,\ \text{Regime}_{f,t}\}$	0.151	0.927	0.932	0.177	0.915	0.935
$\{1,\ \text{Regime}_{f,t},\ d_{t-s}\}$	1.243	0.743	0.770	2.869	0.412	0.505
Panel B: absolute loss						
$\{1,\ \text{Regime}_t\}$	1.443	0.486	0.502	1.416	0.493	0.519
$\{1,\ \text{Regime}_t,\ d_{t-s}\}$	2.733	0.435	0.501	4.499	0.212	0.330
$\{1,\ \text{Regime}_{f,t}\}$	0.896	0.639	0.660	0.740	0.691	0.717
$\{1,\ \text{Regime}_{f,t},\ d_{t-s}\}$	3.099	0.377	0.438	5.484	0.140	0.255

Notes: Wald tests of $\mathbb{E}[d_t \mid \mathcal{F}_t] = 0$ (Giacomini and White 2006). W is the Wald statistic for the null that all coefficients are zero in a regression of the loss difference d_t, the MSC loss minus the professional loss, on the instruments. The covariance is prewhitened Newey–West with a bandwidth of one year, four lags quarterly and two semiannual, and W is asymptotically χ^2_q with $q \in \{2, 3\}$. p_{MBB} is a recentered moving-block bootstrap p-value from 4,999 replications, resampling d_t and the instruments jointly in one-year blocks. Regime_t is the baseline indicator of equation (3), built from smoothed probabilities. $\text{Regime}_{f,t}$ replaces those with filtered probabilities, which condition only on inflation realized through t. It is measurable at the forecast origin and relabels 12% of the quarterly and 14% of the semiannual dates. The instrument d_{t-s} is the most recent loss difference already observable at the origin: a one-year-ahead error is realized four quarters (two half-years) after its own origin, so the latest fully realized loss difference is $s = 5$ quarters (three half-years) old. The augmented samples lose the first s observations, leaving $n = 174$ and 169 quarterly and 87 and 84 semiannual.

voring the professionals, then falls through zero as the volatile quarters of 2021–2023 enter. It never approaches its critical values, 3.01 at the longer window and 3.18 at the shorter, in either direction, and the MSC–Livingston path behaves the same way. Under absolute loss, the paths climb higher, peaking in windows dominated by the stable 2010s, and come within about half a point of their critical values without crossing them. Figure C.1 in Appendix C plots all eight paths, and the appendix reports the statistics. No window in more than four decades delivers a ranking that the full-sample mean conceals.

4.4 Discussion

Our results establish neither a professional accuracy advantage nor a household one. The descriptive gap in Table 2 dissolves under formal testing for three reasons. First, the gaps are small. No regime advantage exceeds 0.20 RMSE points against error levels between 1.01 and 2.26. Second, overlapping twelve-month forecast windows make the errors serially correlated. That persistence cuts the effective size of the stable and volatile quarterly samples by factors of roughly five and six under squared loss.[6] Third, the performance reversal occurs in the noisiest cell of the design. Under squared loss,

[6]The quarterly comparison rests on 110 stable and 64 volatile common dates, and the effective sample sizes fall to 23 and 10, as reported in Table E.1.

differences scale with error magnitude, and the volatile regime makes the errors large precisely where the sample size is smaller. These three are features of the data rather than artifacts of the estimator. No conclusion in Table 3 depends on the estimator choice.

The 95% bootstrap intervals in Table 3 bound the size of any accuracy edge the tests might have missed, measured against the professional surveys' mean loss in Table 2. In the stable regime, the intervals are tight. The largest household advantage consistent with the data is 5% of the SPF's mean absolute error and 16% of its mean squared error, with corresponding bounds of 3% and 13% against the Livingston Survey.[7] In the volatile regime, however, the intervals are wide, admitting household advantages of 25% and 61% of the SPF's mean absolute and mean squared error, and 27% and 65% against the Livingston Survey. The reversal in point estimates therefore sits in the regime where the data offer the least statistical precision.

The limited directional support in the sample favors the professional advantage, and only within the stable regime. In the one-sided bootstrap tail areas of Appendix C, the professional edge in quiet times draws suggestive support, while the household edge in turbulent times draws none. This asymmetry has a practical consequence. Quiet times are precisely when the choice among surveys carries the lowest stakes because errors are small for all three surveys. Turbulent times, when inflation is moving rapidly and expectations are most consequential for policy, are exactly when the choice carries the highest stakes. The presumed professional advantage is thus weakest where a reliable benchmark is needed most.

Because neither survey is demonstrably more accurate, the choice of expectations proxy must be argued rather than presumed. The main question is then not a performance horse race but whether the two groups form expectations the same way. Section 5 addresses that question.

5 Do Households and Professionals Form Expectations Differently?

This section documents two asymmetries between households and professional forecasters. First, the departures from rational expectations that are specific to households occur when inflation is stable. In the volatile regime, inefficiency is robust for the professionals and suggestive for households. Second, when the environment turns

[7] The bound depends on the loss function. Under absolute loss, the stable-regime intervals leave room for a household advantage of at most 5% of professional mean absolute error, while under squared loss the data still admit an advantage of about a sixth of professional mean squared error.

Table 5: Tests of Forecast Bias across Regimes

	MSC	SPF	Livingston
α_1	-0.429	-0.273	-0.260
	(0.264)	(0.219)	(0.173)
	$[-0.878,\ 0.044]$	$[-0.710,\ 0.258]$	$[-0.731,\ 0.290]$
α_2	0.335	0.505	0.528
	(0.780)	(0.686)	(0.554)
	$[-0.417,\ 1.083]$	$[-0.221,\ 1.286]$	$[-0.286,\ 1.372]$
$\alpha_1 + \alpha_2$	-0.094	0.232	0.268
	(0.811)	(0.736)	(0.597)
	$[-0.677,\ 0.509]$	$[-0.334,\ 0.906]$	$[-0.339,\ 0.968]$

Notes: Estimates of equation (6); prewhitened Newey–West standard errors in parentheses. Samples: MSC 519 monthly (1981:M9–2024:M12), SPF 174 quarterly, Livingston 87 semiannual. The October 2024 MSC observation is excluded because its twelve-month-ahead realization (the October 2025 CPI) was never published owing to the 2025 U.S. federal government shutdown. Brackets: residual moving-block bootstrap 95% percentile intervals from 4,999 replications.

volatile, the two groups move the weight on recently realized inflation in opposite directions: households toward it, professionals away.

Section 5.1 runs the four rationality tests behind the first asymmetry, Section 5.2 estimates the adaptive-expectations regressions behind the second, and Section 5.3 discusses what the two asymmetries mean. Appendix D supports the section with three further exercises: we re-estimate every specification without regime terms in Tables D.1 and D.2, replace the distributed lag with a single inflation lag in Table D.3, and estimate a revision-based alternative to the adaptive regression in Table D.4.

5.1 Four dimensions of rationality

Bias. No survey shows statistically significant regime-conditional bias in Table 5, where we estimate equation (6). The stable-regime point estimates of α_1 are negative for all three surveys, at -0.429 for the MSC, -0.273 for the SPF, and -0.260 for Livingston, but none is significant. The volatile-regime sums are small, of mixed sign, and insignificant. We read the estimates as suggesting mild stable-regime over-prediction for both groups without establishing it.

Own-forecast efficiency. Under informational efficiency the forecast error should be orthogonal to the forecast itself. Table 6, where we estimate equation (7), supports one summary: volatile-regime own-forecast inefficiency is robustly established for the professionals. For households, the point estimate is as large but the evidence is route-sensitive.

The professional half is robust. In the stable regime, β_1 is insignificant for both

Table 6: Tests of Own-Forecast Efficiency across Regimes

	MSC	SPF	Livingston
β_1	-0.208 (0.479) $[-1.018,\ 0.614]$	-0.230 (0.272) $[-0.681,\ 0.274]$	-0.173 (0.222) $[-0.641,\ 0.335]$
β_2	-0.706 (0.809) $[-1.672,\ 0.270]$	-0.601^{*} (0.345) $[-1.196,\ -0.030]$	-0.698^{**} (0.289) $[-1.324,\ -0.064]$
$\beta_1 + \beta_2$	-0.913 (0.736) $[-1.461,\ -0.341]$	-0.831^{***} (0.288) $[-1.223,\ -0.415]$	-0.871^{***} (0.245) $[-1.285,\ -0.399]$
α_1	0.179 (1.215) $[-2.275,\ 2.588]$	0.368 (0.809) $[-1.055,\ 1.774]$	0.228 (0.634) $[-1.250,\ 1.647]$
α_2	3.014 (2.157) $[-0.185,\ 6.101]$	2.565^{*} (1.543) $[0.682,\ 4.490]$	2.742^{**} (1.299) $[0.707,\ 4.796]$
$\alpha_1 + \alpha_2$	3.193^{*} (1.897) $[1.104,\ 5.232]$	2.933^{*} (1.538) $[1.526,\ 4.419]$	2.969^{**} (1.300) $[1.506,\ 4.508]$
L1: Reject H_0: $\alpha_1 = \beta_1 = 0$?	F=1.322 p=0.268 p_{MBB}=0.526	F=1.287 p=0.279 p_{MBB}=0.503	F=1.376 p=0.258 p_{MBB}=0.512
L2: Reject H_0: $\alpha_1 + \alpha_2 = \beta_1 + \beta_2 = 0$?	F=2.062 p=0.128 p_{MBB}=0.230	F=17.103 p<0.001 p_{MBB}=0.003	F=29.012 p<0.001 p_{MBB}=0.001

Notes: Estimates of equation (7); prewhitened Newey–West standard errors in parentheses. Brackets: residual moving-block bootstrap 95% percentile intervals from 4,999 replications; p_{MBB} beneath L1 and L2: null-imposed residual-MBB p-values from 1,999 replications. Samples as in Table 5. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

surveys. In the volatile regime, the sum $\beta_1 + \beta_2$ is significantly negative, at -0.831 for the SPF and -0.871 for Livingston, both at the 1% level, and the joint test rejects overwhelmingly, with F statistics of 17.1 and 29.0. A coefficient of this size is economically large. In volatile times, a forecast one percentage point higher comes with over-prediction nearly nine tenths of a point larger, predictable from the forecast alone.

The household half is weaker: the stable-regime coefficient of -0.208 and the volatile-regime sum of -0.913 are both insignificant under the prewhitened standard errors, and neither joint test rejects. The volatile-regime reading is route-sensitive: the bootstrap interval for the sum excludes zero while the null-imposed joint bootstrap does not reject, so we treat the household evidence as suggestive rather than established.

Error persistence. Only household errors fail equation (8), and only in the stable regime, as reported in Table 7. When inflation is stable, household errors show significant positive persistence, with $\beta_1 = 0.334$ at the 5% level, though the moving-block bootstrap interval of $[-0.050,\ 0.707]$ narrowly includes zero, so we read the persistence as suggestive rather than established. The persistence vanishes in the volatile regime,

Table 7: Tests of Forecast-Error Persistence across Regimes

	MSC	SPF	Livingston
β_1	0.334^{**}	0.120	0.169
	(0.168)	(0.181)	(0.203)
	$[-0.050, 0.707]$	$[-0.394, 0.644]$	$[-0.459, 0.805]$
β_2	-0.422	0.027	0.018
	(0.317)	(0.293)	(0.283)
	$[-0.857, 0.020]$	$[-0.537, 0.588]$	$[-0.661, 0.704]$
$\beta_1 + \beta_2$	-0.088	0.147	0.186
	(0.288)	(0.214)	(0.173)
	$[-0.334, 0.156]$	$[-0.082, 0.385]$	$[-0.068, 0.450]$
α_1	-0.270	-0.231	-0.204
	(0.211)	(0.189)	(0.158)
	$[-0.741, 0.239]$	$[-0.662, 0.279]$	$[-0.692, 0.358]$
α_2	0.176	0.643	0.417
	(0.713)	(0.464)	(0.422)
	$[-0.558, 0.906]$	$[-0.120, 1.425]$	$[-0.384, 1.257]$
$\alpha_1 + \alpha_2$	-0.094	0.411	0.213
	(0.748)	(0.479)	(0.418)
	$[-0.659, 0.489]$	$[-0.127, 1.091]$	$[-0.365, 0.900]$

Notes: Estimates of equation (8); prewhitened Newey–West standard errors in parentheses. Samples as in Table 5, except that a forecast dated twelve months earlier is additionally required, leaving 170 SPF and 87 Livingston observations. Brackets: residual moving-block bootstrap 95% percentile intervals from 4,999 replications. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

and professional errors show none in either regime. Stable-regime persistence is what sluggish belief updating looks like when the cost of inattention is low.

Use of macroeconomic information. The two groups fail the macroeconomic-information restriction in mirror image: household errors are predictable from public macroeconomic data when inflation is stable, professional errors when it is volatile. Rationality requires that lagged inflation, the short-term interest rate, and the unemployment rate not predict the forecast error. We estimate equation (9) in Table 8, with L3 testing the stable-regime coefficients and L4 their volatile-regime sums.

For households, L3 rejects at the 5% level, with $F = 2.732$ and $p = 0.043$, while the volatile-regime test does not reject. The interest-rate coefficient $\kappa_1 = 0.206$ is positive and significant at the 10% level, so inflation later runs above the household forecast when interest rates are high, which points to under-reaction to interest-rate news. For the professionals, the pattern runs the other way. L3 does not reject, while L4 rejects at $p = 0.038$ for the SPF and $p = 0.078$ for Livingston. The cumulative unemployment coefficient is significantly negative for both.

One contrast invites misreading. The MSC stable-regime $\beta_1 = -1.026$, with a standard error of 0.371, significant at the 1% level, does not contradict the insignificant -0.208 of Table 6. Conditioning on the lagged block of π_{t-1}, i_{t-1}, and $\mathcal{U}_{t-1}$ makes β_1

the loading on the forecast component orthogonal to those variables, an object best read through the joint L3 test rather than the single coefficient.

Both halves of the mirror-image pattern need qualification, because neither survives the bootstrap route of Table 8. The household stable-regime rejection carries a null-imposed moving-block bootstrap p-value of 0.089, so we read it as suggestive rather than established and do not build on it. The professional volatile-regime rejections fare no better, with bootstrap p-values of 0.214 for the SPF and 0.390 for Livingston, and we read them as suggestive as well.

The rationality asymmetry. Tables 5–8 together show an asymmetry in where rationality fails. The household-specific departures occur in the stable regime. They are persistent errors and unexploited macroeconomic information, both only suggestive. Volatile-regime own-forecast inefficiency is robust for the professionals and only suggestive for households. Conditioning makes the asymmetry visible. Without regime interactions the tests collapse to familiar full-sample results: efficiency rejected for both professional surveys but not the MSC, macroeconomic information rejected for no survey, as reported in Table D.1 in Appendix D.

5.2 Expectation adaptiveness

When the inflation environment turns volatile, households shift their forecasts toward recently realized inflation and professionals shift away from it. In Table 9, we estimate the distributed-lag regression of equation (10) and report the cumulative loadings defined there, $\beta_1(1)$ for the stable regime and $\beta_1(1) + \beta_2(1)$ for the volatile regime. Three results follow.

First, in the stable regime the three surveys are nearly identical. The cumulative loadings are 0.305 for the MSC, 0.284 for the SPF, and 0.302 for Livingston, each significant at the 1% level and within a few hundredths of the others. Whatever separates households from professionals is not visible when inflation is quiet.

Second, in the volatile regime they diverge in opposite directions. The regime interaction $\beta_2(1)$ is positive and significant for households, at 0.079 with $p < 0.01$. For the SPF, it is negative though insignificant, at -0.042, and for Livingston it is negative and significant, at -0.109 with $p < 0.01$. Cumulatively, the household loading rises to 0.383 while the professional loadings fall to 0.242 for the SPF and 0.193 for Livingston.[8] Household expectations now move with realized inflation between roughly 1.6 and 2

[8] Allowing the intercept to vary by regime gives volatile-regime loadings of 0.383, 0.285, and 0.225, each significant at the 1% level, with the ratio then between 1.3 and 1.7.

Table 8: Tests of the Use of Macroeconomic Information across Regimes

	MSC	SPF	Livingston
α_1	0.710	0.283	−0.058
	(1.085)	(0.944)	(0.908)
	[−1.696, 3.120]	[−1.519, 2.134]	[−1.891, 1.755]
α_2	3.595**	3.361**	3.642**
	(1.588)	(1.381)	(1.489)
	[0.050, 7.259]	[0.558, 6.156]	[0.586, 6.684]
β_1	−1.026***	−0.550	−0.674
	(0.371)	(0.540)	(0.491)
	[−1.889, −0.175]	[−1.512, 0.486]	[−1.757, 0.417]
β_2	−0.279	1.004	0.726
	(0.875)	(0.980)	(0.833)
	[−1.293, 0.767]	[−0.408, 2.446]	[−0.750, 2.228]
γ_1	0.133	0.038	0.000
	(0.232)	(0.179)	(0.188)
	[−0.393, 0.641]	[−0.490, 0.570]	[−0.583, 0.603]
γ_2	0.045	−0.347	−0.173
	(0.311)	(0.236)	(0.214)
	[−0.546, 0.662]	[−0.990, 0.289]	[−0.836, 0.500]
$\gamma_1 + \gamma_2$	0.178	−0.309*	−0.172
	(0.187)	(0.180)	(0.126)
	[−0.128, 0.490]	[−0.704, 0.063]	[−0.544, 0.196]
κ_1	0.206*	0.103	0.179
	(0.112)	(0.172)	(0.163)
	[−0.026, 0.446]	[−0.210, 0.427]	[−0.165, 0.523]
κ_2	−0.215	−0.473	−0.471
	(0.256)	(0.422)	(0.398)
	[−0.506, 0.067]	[−0.910, −0.050]	[−0.962, 0.009]
$\kappa_1 + \kappa_2$	−0.009	−0.370	−0.291
	(0.222)	(0.335)	(0.303)
	[−0.199, 0.162]	[−0.701, −0.051]	[−0.634, 0.059]
δ_1	0.142	0.093	0.187
	(0.187)	(0.202)	(0.178)
	[−0.107, 0.403]	[−0.187, 0.354]	[−0.095, 0.467]
δ_2	−0.192	−0.414	−0.439*
	(0.217)	(0.275)	(0.248)
	[−0.577, 0.198]	[−0.828, 0.023]	[−0.884, 0.010]
$\delta_1 + \delta_2$	−0.050	−0.322**	−0.251**
	(0.164)	(0.133)	(0.120)
	[−0.349, 0.259]	[−0.648, 0.024]	[−0.587, 0.096]
L3: Reject H_0: $\gamma_1 = \kappa_1 = \delta_1 = 0$?	F=2.732	F=0.326	F=0.642
	p=0.043	p=0.807	p=0.590
	p_{MBB}=0.089	p_{MBB}=0.869	p_{MBB}=0.745
L4: Reject H_0: $\gamma_1 + \gamma_2 = \kappa_1 + \kappa_2 = \delta_1 + \delta_2 = 0$?	F=0.322	F=2.865	F=2.359
	p=0.809	p=0.038	p=0.078
	p_{MBB}=0.861	p_{MBB}=0.214	p_{MBB}=0.390

Notes: Estimates of equation (9); prewhitened Newey–West standard errors in parentheses. Brackets: residual moving-block bootstrap 95% percentile intervals from 4,999 replications; p_{MBB} beneath L3 and L4: null-imposed residual-MBB p-values from 1,999 replications. Samples: MSC 519 monthly (1981:M9–2024:M12), SPF 174 quarterly, Livingston 87 semiannual. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

times as strongly as professional expectations, whereas in the stable regime the ratio is one to one.

This opposite-signed divergence is the paper's central finding, and it does not rest on the Newey–West route alone. The residual moving-block bootstrap agrees survey by survey. The household shift interval, $[0.033,\ 0.125]$, sits above zero, the SPF interval, $[-0.102,\ 0.019]$, covers zero, and the Livingston interval, $[-0.175,\ -0.040]$, sits below zero. The same sign pattern reappears in the single-lag specification of Table D.3 in Appendix D.

Third, none of the surveys is purely backward-looking. The macroeconomic controls are jointly significant for all three surveys, with unemployment terms individually significant for households and interest-rate terms for the professionals. The divergence is therefore relative, not absolute. Unconditionally, when Table D.2 of Appendix D re-estimates the regression without regime terms, the household cumulative loading is 0.404, 1.8 times the 0.225 of the SPF and 2.4 times the 0.169 of Livingston, and the macroeconomic controls remain jointly significant.

We further check that the divergence is specific to the margin the adaptive regression measures. Following Pfajfar and Santoro (2010), we regress the forecast revision on the most recent verifiable forecast error, a margin that measures how fast each survey corrects its own mistakes, and report the estimates in Table D.4 of Appendix D. That margin shows no regime shift: the interaction is insignificant for all three surveys, though household revisions respond at least as strongly as professional revisions. What moves across regimes is therefore the adaptiveness of equation (10), the weight expectations place on recently realized inflation, not the speed at which errors are corrected.

5.3 Discussion

When inflation is stable, the choice of survey proxy makes little difference to the weight expectations place on recent inflation. All three measures are close to interchangeable there, and none of them shows significant bias or own-forecast inefficiency. When the environment turns volatile, however, the two groups form expectations differently. Households raise the pass-through of recently realized inflation while professional forecasters lower it. This divergence emerges precisely when expectations matter most for wage bargaining, durable goods purchases, and monetary policy.

The departures from rationality follow the same divide. In the volatile regime, the two groups' errors become predictable from their own forecasts. In the stable regime, where the cost of inattention is low, household errors show persistence and

Table 9: Tests of Adaptive Expectations across Regimes

	MSC	SPF	Livingston
$\beta_1(1)$	0.305^{***} (0.038) [0.227, 0.382]	0.284^{***} (0.070) [0.185, 0.388]	0.302^{***} (0.060) [0.185, 0.410]
$\beta_2(1)$	0.079^{***} (0.030) [0.033, 0.125]	-0.042 (0.040) $[-0.102, 0.019]$	-0.109^{***} (0.031) $[-0.175, -0.040]$
$\beta_1(1) + \beta_2(1)$	0.383^{***} (0.019) [0.329, 0.437]	0.242^{***} (0.040) [0.168, 0.317]	0.193^{***} (0.044) [0.112, 0.271]
α	2.059^{***} (0.150) [1.760, 2.341]	0.329 (0.325) $[-0.072, 0.731]$	0.278 (0.309) $[-0.143, 0.670]$
λ_1	-0.078^{***} (0.019) $[-0.145, -0.008]$	0.118 (0.156) $[-0.166, 0.402]$	0.013 (0.144) $[-0.373, 0.392]$
λ_2	0.115^{***} (0.023) [0.045, 0.182]	0.039 (0.138) $[-0.225, 0.304]$	0.158 (0.150) $[-0.203, 0.525]$
θ_1	-0.046 (0.117) $[-0.202, 0.116]$	0.321^{**} (0.143) [0.046, 0.584]	0.727^{**} (0.298) [0.192, 1.258]
θ_2	0.009 (0.111) $[-0.153, 0.166]$	-0.082 (0.136) $[-0.347, 0.195]$	-0.494^{*} (0.286) $[-1.006, 0.029]$
L5: Reject H_0: $\lambda_1 = \lambda_2 = \theta_1 = \theta_2 = 0$?	F=9.261 p<0.001 p_{MBB}=0.020	F=31.026 p<0.001 p_{MBB}=0.001	F=22.023 p<0.001 p_{MBB}=0.002

Notes: Estimates of equation (10); prewhitened Newey–West standard errors in parentheses. α is the intercept, which does not vary by regime. The controls λ_1 and λ_2 multiply current and lagged unemployment, θ_1 and θ_2 current and lagged interest rates. p_{MBB} beneath L5: its null-imposed residual-MBB p-value from 1,999 replications. Brackets: residual moving-block bootstrap 95% percentile intervals from 4,999 replications. Samples: MSC 520 monthly (1981:M9–2024:M12; October 2024 re-enters because the regression requires no twelve-month-ahead realization), SPF 174 quarterly, Livingston 87 semiannual. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

are predictable from public macroeconomic data. Of these, only the professionals' own-forecast inefficiency is robust, and the household results are suggestive.

Read together with Section 4, these results form the paper's central pattern. Neither group forecasts demonstrably better, yet the two groups respond in opposite ways when inflation turns volatile. What separates households from professionals is not how well they forecast but how they form the forecast.

The pattern reads naturally as state-dependent attention. When inflation turns volatile, the return to tracking it rises for households, and their expectations lean harder on recent realizations, while professionals, who monitor in every environment, shift weight the other way. We do not observe information acquisition directly, so we offer this as an interpretation rather than as causal inference.

In volatile times, the two surveys are not interchangeable proxies for one latent expectation but different signals. This difference has two implications. For central banks, reading professional forecasts alone would understate how much recently realized inflation is entering household expectations, so the two series are better monitored together. The gap between them is itself informative. When it widens, recent inflation is passing into household expectations faster than into professional ones. For applied work, a Phillips curve or a monetary-policy rule estimated with household expectations embeds a different formation process than one estimated with professional forecasts, and forecast accuracy cannot decide between the two. Which measure to use is therefore a modeling assumption to make explicit.

6 Robustness Checks

The preceding results rest on four choices: the survey frequency, the regime definition, the standard-error estimator, and the treatment of the estimated regime's first-stage uncertainty. This section addresses each in turn. The checks target the main results the paper's conclusions rest on: the loss-difference tests under squared loss, the professional own-forecast efficiency in both regimes, and the cumulative loadings $\beta_1(1)$ and $\beta_1(1) + \beta_2(1)$. All three survive every check that bears on them.

Section 6.1 re-estimates the household loadings on the professional surveys' observation dates. Section 6.2 re-runs the core results under six regime classifications. Section 6.3 compares the prewhitened estimator with its alternatives, and Section 6.4 propagates the first-stage uncertainty of the estimated regime into the results. Appendix E collects the supporting tables and the bootstrap procedures.

Table 10: Common-Date Robustness of the Household Loadings

Result	MSC at SPF dates (quarterly)	MSC at Livingston dates (semiannual)
Adaptive loading, stable ($\beta_1(1)$)	0.303 ($p < 0.001$)	0.295 ($p < 0.001$)
Adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.368 ($p < 0.001$)	0.342 ($p < 0.001$)

Notes: The frequency check of Section 6.1: the MSC cumulative loadings of the Table 9 specification, re-estimated on the professional surveys' observation dates, 174 quarterly and 87 semiannual, with the prewhitened Newey–West bandwidth reset to one year at each frequency, four and two lags. Cells report the loading with the p-value in parentheses.

6.1 Frequency and common observation dates

Two of the three target results need no frequency check. We compute the loss differences on the common observation dates by construction, so Table 3 already is the common-date result, and we estimate the professional efficiency and loading results at the professionals' own dates throughout. The check therefore bears on the household side. Table 10 re-estimates the MSC cumulative loadings on the professional surveys' observation dates, 174 quarterly and 87 semiannual, with the HAC bandwidth reset to one year at each frequency.

The loadings survive. $\beta_1(1)$ is 0.303 quarterly and 0.295 semiannual, the volatile-regime sum $\beta_1(1) + \beta_2(1)$ is 0.368 and 0.342, and all four are significant at the 1% level.

6.2 Alternative regime definitions

The baseline indicator of equation (3) is an *ex post* description of the inflation environment: smoothed probabilities, which condition on the full sample, and a cutoff of one-half. An *ex post* description is the right object for our question, which asks how expectations behaved in environments that were in fact turbulent, but the indicator embeds choices, and the smoothed probability at t draws on inflation realized after t. The one test where the reading is not ours to choose is the conditional predictive ability test of Section 4.3, whose instruments must be measurable at the forecast origin. We therefore consider six classifications, the baseline plus five variants: the continuous smoothed probability in place of the binary indicator; filtered rather than smoothed probabilities, which use only information up to t and so remove the look-ahead; cutoffs of $c \in \{0.4, 0.5, 0.6\}$, the middle one the baseline; and the filtered indicator dated one quarter before the forecast origin. The filtered probability at t still conditions on inflation realized during the origin quarter, which the expectations formed at t both observe and can feed into. Dating the indicator one quarter earlier removes the look-ahead and this contemporaneous feedback at once, and the resulting variant is deliberately the noisiest classification. The cutoff variants, by contrast, can differ only at the

margin: seven of the 174 evaluation-window quarters carry a smoothed probability between 0.4 and 0.6, so the three cutoffs draw nearly the same line.

Tables 11 and 12 report each core quantity under each classification. The professional efficiency contrast is the most robust result. The volatile-regime sum is significant at 5% or better under every classification for both surveys, including the lagged filtered indicator, where the sums are -0.705 for the SPF and -0.702 for Livingston. The stable-regime coefficient β_1 stays insignificant under every classification, with the smallest p-value at 0.33, so the inefficiency appears in the volatile regime and only there.

The volatile-regime cumulative loadings barely move across classifications. The household loading stays between 0.371 and 0.383, the SPF loading between 0.232 and 0.246, and the Livingston loading between 0.171 and 0.196, every entry significant at the 1% level with a residual-MBB interval strictly above zero. The household weight on recent inflation is therefore between roughly 1.5 and 2.2 times the professional weight under every classification, including the deliberately noisy lagged filtered indicator. The stable-regime loadings stay between 0.250 and 0.334 across surveys and classifications, each significant at the 1% level. The interaction terms behind the sums behave as at the baseline: the household shift is positive and significant under the continuous, cutoff, and filtered classifications, the Livingston shift is negative and significant under all six, and the SPF shift is negative but insignificant throughout. The one attenuation is the household shift under the lagged filtered indicator, at an insignificant 0.037 with a standard error of 0.030. The same column shows why: the volatile-regime loss differences collapse toward zero under this variant, to -0.008 and 0.198 from -0.592 and -0.588 at the baseline, so the doubly lagged classification no longer separates turbulent from tranquil environments. Attenuation toward zero is what misclassification of a binary conditioning variable produces, and the loading levels hold steady through it.

On the accuracy side, no mean loss difference is significant under any classification, in either regime. Signs behave as the baseline does, with two exceptions, both in the lagged filtered MSC–Livingston column: there the volatile-regime difference turns slightly positive and the stable-regime difference slightly negative.

The filtered-probability column is the closest any variant comes to real-time construction with contemporaneous data. The lagged filtered column goes one step further and prices the residual noise. The professional volatile-regime inefficiency survives both variants. Under the filtered indicator, the sums are -0.703 for the SPF and -0.721 for Livingston, both significant at the 1% level. The household adaptive-loading shift survives the filtered variant, at 0.066 and significant at the 5% level, but

Table 11: Regime-Definition Robustness: Accuracy

Result	Continuous probability	Filtered probability	Filtered, lagged one quarter
MSC−SPF loss difference, stable	0.602 (0.594) [−0.151, 1.351]	0.150 (0.446) [−0.541, 0.792]	0.098 (0.522) [−0.806, 0.848]
MSC−SPF loss difference, volatile	−0.904 (2.818) [−4.122, 2.656]	−0.190 (2.181) [−2.752, 2.815]	−0.008 (1.567) [−1.960, 2.801]
MSC−Livingston loss difference, stable	0.510 (0.448) [−0.212, 1.236]	0.134 (0.353) [−0.406, 0.675]	−0.085 (0.588) [−1.223, 0.773]
MSC−Livingston loss difference, volatile	−0.843 (2.033) [−4.444, 2.983]	−0.252 (1.749) [−3.408, 3.065]	0.198 (1.050) [−1.632, 2.866]

Result	$c = 0.4$	$c = 0.5$ (baseline)	$c = 0.6$
MSC−SPF loss difference, stable	0.494 (0.380) [−0.049, 1.072]	0.405 (0.410) [−0.173, 0.993]	0.345 (0.456) [−0.266, 0.958]
MSC−SPF loss difference, volatile	−0.655 (2.063) [−3.027, 2.020]	−0.592 (2.185) [−3.120, 2.233]	−0.516 (2.174) [−3.113, 2.334]
MSC−Livingston loss difference, stable	0.426 (0.330) [−0.020, 0.929]	0.357 (0.321) [−0.131, 0.885]	0.268 (0.381) [−0.253, 0.823]
MSC−Livingston loss difference, volatile	−0.641 (1.498) [−3.241, 2.208]	−0.588 (1.538) [−3.327, 2.397]	−0.481 (1.559) [−3.430, 2.638]

Notes: The regime-definition check of Section 6.2, accuracy side: mean loss differences $\bar{d}$ under squared loss by regime, negative entries indicating lower household loss; no entry is significantly different from zero. Each cell reports the estimate, its prewhitened Newey–West standard error in parentheses, and a moving-block bootstrap 95% interval in brackets from 9,999 replications, resampling the loss difference and the classification jointly. The baseline column ($c = 0.5$, smoothed probabilities) corresponds to the main-text tables; the filtered-probability column uses the 0.5 cutoff on filtered probabilities; the lagged filtered column dates that indicator one quarter before the forecast origin, so the first three window months drop out; under the continuous specification we replace the regime dummy with the smoothed probability p_t and evaluate the volatile-regime mean at $p_t = 1$ and the stable-regime mean at $p_t = 0$.

attenuates under the lagged one, as recorded above. Any operational reading should therefore rest on the professional inefficiency and the household loading shift.

6.3 HAC estimator and block-bootstrap inference

The serial correlation that the twelve-month overlap builds into the errors is strong, with a first-order autocorrelation of 0.74 in the quarterly loss difference, and kernel-only HAC estimators over-reject when correlation is this strong. The baseline of Section 2.3 therefore prewhitens: a first-order vector autoregressive step precedes the Bartlett kernel. Prewhitening removes most of the size distortion, and it is also the conservative choice here, raising the standard error relative to the non-prewhitened estimator in nearly every series we examine. This subsection checks the baseline four ways.

Table 12: Regime-Definition Robustness: Rationality and Formation

Result	Continuous probability	Filtered probability	Filtered, lagged one quarter
SPF own-forecast efficiency, stable (β_1)	-0.153	-0.325	-0.350
	(0.334)	(0.380)	(0.415)
	$[-0.712,\ 0.457]$	$[-0.794,\ 0.226]$	$[-0.829,\ 0.180]$
Livingston own-forecast efficiency, stable (β_1)	-0.116	-0.291	-0.338
	(0.284)	(0.347)	(0.377)
	$[-0.715,\ 0.522]$	$[-0.802,\ 0.282]$	$[-0.827,\ 0.210]$
SPF own-forecast efficiency, volatile ($\beta_1 + \beta_2$)	-0.890^{**}	-0.703^{***}	-0.705^{***}
	(0.352)	(0.230)	(0.193)
	$[-1.324,\ -0.443]$	$[-1.098,\ -0.244]$	$[-1.138,\ -0.215]$
Livingston own-forecast efficiency, volatile ($\beta_1 + \beta_2$)	-0.924^{***}	-0.721^{***}	-0.702^{***}
	(0.284)	(0.198)	(0.117)
	$[-1.393,\ -0.361]$	$[-1.138,\ -0.208]$	$[-1.134,\ -0.157]$
MSC adaptive loading, stable ($\beta_1(1)$)	0.250^{***}	0.317^{***}	0.334^{***}
	(0.043)	(0.045)	(0.043)
	$[0.156,\ 0.345]$	$[0.229,\ 0.403]$	$[0.246,\ 0.419]$
SPF adaptive loading, stable ($\beta_1(1)$)	0.285^{***}	0.288^{***}	0.288^{***}
	(0.094)	(0.073)	(0.073)
	$[0.160,\ 0.414]$	$[0.177,\ 0.399]$	$[0.176,\ 0.394]$
Livingston adaptive loading, stable ($\beta_1(1)$)	0.265^{***}	0.262^{***}	0.271^{***}
	(0.079)	(0.064)	(0.073)
	$[0.117,\ 0.406]$	$[0.144,\ 0.375]$	$[0.154,\ 0.384]$
MSC adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.373^{***}	0.383^{***}	0.371^{***}
	(0.019)	(0.023)	(0.025)
	$[0.317,\ 0.428]$	$[0.325,\ 0.440]$	$[0.311,\ 0.430]$
SPF adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.242^{***}	0.244^{***}	0.232^{***}
	(0.046)	(0.043)	(0.042)
	$[0.167,\ 0.318]$	$[0.168,\ 0.320]$	$[0.158,\ 0.310]$
Livingston adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.171^{***}	0.182^{***}	0.189^{***}
	(0.049)	(0.044)	(0.047)
	$[0.089,\ 0.252]$	$[0.103,\ 0.261]$	$[0.107,\ 0.269]$

Notes: The table continues on the next page with the cutoff classifications, where the full notes appear. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table 12: Regime-Definition Robustness: Rationality and Formation (continued)

Result	$c = 0.4$	$c = 0.5$ (baseline)	$c = 0.6$
SPF own-forecast efficiency, stable (β_1)	-0.185	-0.230	-0.299
	(0.232)	(0.272)	(0.330)
	$[-0.624, 0.310]$	$[-0.681, 0.274]$	$[-0.750, 0.222]$
Livingston own-forecast efficiency, stable (β_1)	-0.185	-0.173	-0.292
	(0.212)	(0.222)	(0.296)
	$[-0.664, 0.320]$	$[-0.641, 0.335]$	$[-0.757, 0.235]$
SPF own-forecast efficiency, volatile ($\beta_1 + \beta_2$)	-0.846^{***}	-0.831^{***}	-0.787^{***}
	(0.324)	(0.288)	(0.251)
	$[-1.223, -0.460]$	$[-1.223, -0.415]$	$[-1.178, -0.355]$
Livingston own-forecast efficiency, volatile ($\beta_1 + \beta_2$)	-0.868^{***}	-0.871^{***}	-0.781^{***}
	(0.278)	(0.245)	(0.188)
	$[-1.267, -0.423]$	$[-1.285, -0.399]$	$[-1.211, -0.258]$
MSC adaptive loading, stable ($\beta_1(1)$)	0.309^{***}	0.305^{***}	0.303^{***}
	(0.040)	(0.038)	(0.039)
	$[0.228, 0.390]$	$[0.227, 0.382]$	$[0.224, 0.380]$
SPF adaptive loading, stable ($\beta_1(1)$)	0.301^{***}	0.284^{***}	0.284^{***}
	(0.071)	(0.070)	(0.071)
	$[0.197, 0.406]$	$[0.185, 0.388]$	$[0.185, 0.388]$
Livingston adaptive loading, stable ($\beta_1(1)$)	0.293^{***}	0.302^{***}	0.304^{***}
	(0.060)	(0.060)	(0.066)
	$[0.173, 0.410]$	$[0.185, 0.410]$	$[0.188, 0.415]$
MSC adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.383^{***}	0.383^{***}	0.382^{***}
	(0.020)	(0.019)	(0.019)
	$[0.329, 0.437]$	$[0.329, 0.437]$	$[0.328, 0.436]$
SPF adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.246^{***}	0.242^{***}	0.242^{***}
	(0.042)	(0.040)	(0.040)
	$[0.173, 0.320]$	$[0.168, 0.317]$	$[0.169, 0.317]$
Livingston adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.187^{***}	0.193^{***}	0.196^{***}
	(0.044)	(0.044)	(0.045)
	$[0.107, 0.267]$	$[0.112, 0.271]$	$[0.115, 0.275]$

Notes: The regime-definition check of Section 6.2, formation side: own-forecast efficiency rows report the stable-regime coefficient β_1 and the volatile-regime sum $\beta_1 + \beta_2$, adaptive rows the cumulative loadings $\beta_1(1)$ and $\beta_1(1) + \beta_2(1)$. Each cell reports the estimate, its prewhitened Newey–West standard error in parentheses, and a residual-MBB 95% interval in brackets from 4,999 replications. The baseline column ($c = 0.5$, smoothed probabilities) corresponds to the main-text tables; the filtered-probability column uses the 0.5 cutoff on filtered probabilities; the lagged filtered column dates that indicator one quarter before the forecast origin, so the first three window months drop out; under the continuous specification we replace the regime dummy with the smoothed probability p_t, evaluating volatile-regime quantities at $p_t = 1$ and stable-regime quantities at $p_t = 0$. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Serial correlation. In the loss-difference autocorrelations of Table E.1 in Appendix E, the first q are large and significant in every series, as the twelve-month overlap predicts, and beyond lag q a joint test on lags $q+1$ through $q+4$ fails to reject in both squared-loss series. The fixed one-year bandwidth and block length are therefore adequate, but the serial correlation is costly: the variance inflation factor reaches 6.7 for the quarterly squared-loss difference.

Prewhitening. We recompute every main-table standard error and joint test with the non-prewhitened Newey–West estimator, point estimates fixed. On the three targets, the loss differences stay insignificant, the professional efficiency contrast of Table 6 is intact, with the volatile-regime rejections stronger and the stable-regime coefficients still insignificant, and all of Table 9 is robust.

Small-sample corrections. The baseline takes p-values from a t distribution but applies neither the $n/(n-k)$ covariance scaling nor the Harvey, Leybourne, and Newbold (1997) (HLN) correction to the Diebold–Mariano statistic. Table E.2 adds both to every loss-difference cell, and neither changes a conclusion. The HLN correction lowers the p-value in most cells and raises it in one by 0.018, so the paper's convention is generally the more conservative. The smallest p-value under any combination is 0.224.

Moving-block bootstrap. The moving-block bootstrap complements we report throughout the tables, the bracketed intervals and p_{MBB} entries, are a second inference route that bypasses HAC asymptotics. Section 2.3 introduces them and Appendix E.3 gives the mechanics. The route confirms the headline quantities, collected in Table E.3: the professional volatile-regime inefficiency intervals exclude zero, at $[-1.223,\ -0.415]$ for the SPF and $[-1.285,\ -0.399]$ for Livingston, as do the six cumulative-loading intervals, stable and volatile for all three surveys, and the household loading-shift interval, $[0.033,\ 0.125]$. The stable-regime efficiency intervals straddle zero, at $[-0.681,\ 0.274]$ for the SPF and $[-0.641,\ 0.335]$ for Livingston, in line with Table 6.

6.4 Regime-estimation uncertainty

The regime indicator is a generated regressor: we estimate it in a first stage, and second-stage standard errors that treat it as known omit the first-stage uncertainty. We quantify the omission with a parametric bootstrap of the Markov-switching model, parametric because each replication simulates a fresh inflation history from the estimated MS-AR(1) rather than resampling blocks of data. Each replication re-estimates the model on its simulated history, uses the re-estimated parameters to reclassify the actual series at the baseline cutoff, and re-runs the second stage. Holding the data fixed and varying only the classification isolates the one channel through which first-stage

Table 13: Parametric MS Bootstrap: Core Quantities

Result	Baseline estimate	Bootstrap 95% interval
MSC−SPF loss difference, stable	0.405	[0.073, 0.630]
MSC−SPF loss difference, volatile	−0.592	[−1.185, 0.031]
MSC−Livingston loss difference, stable	0.357	[0.018, 0.577]
MSC−Livingston loss difference, volatile	−0.588	[−1.037, 0.009]
SPF own-forecast efficiency, stable (β_1)	−0.230	[−0.663, −0.088]
Livingston own-forecast efficiency, stable (β_1)	−0.173	[−0.618, −0.036]
SPF own-forecast efficiency, volatile ($\beta_1 + \beta_2$)	−0.831	[−0.897, −0.573]
Livingston own-forecast efficiency, volatile ($\beta_1 + \beta_2$)	−0.871	[−0.930, −0.524]
MSC adaptive loading, stable ($\beta_1(1)$)	0.305	[0.258, 0.389]
SPF adaptive loading, stable ($\beta_1(1)$)	0.284	[0.176, 0.315]
Livingston adaptive loading, stable ($\beta_1(1)$)	0.302	[0.135, 0.309]
MSC adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.383	[0.371, 0.432]
SPF adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.242	[0.198, 0.254]
Livingston adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.193	[0.118, 0.206]

Notes: The regime-uncertainty check of Section 6.4: baseline estimates from Tables 3, 6, and 9, with the loss differences under squared loss. Each interval reports the 2.5–97.5 percentiles across 999 parametric-bootstrap replications of the first-stage MS-AR(1), all successful, each entry using the 978–999 replications for which the quantity is defined; the data are held fixed and only the regime classification varies, following the four-step procedure of Appendix E.4, so the intervals reflect first-stage regime-estimation uncertainty only. The median replication retains the baseline classification for 93.1% of evaluation-window months. The volatile-regime loss differences are negative in 97.3% (SPF) and 97.2% (Livingston) of replications and the stable-regime ones positive in 98.2% and 97.8%; the volatile-regime household loading exceeds the SPF loading in 100.0% and the Livingston loading in 99.9% of replications, and the volatile-regime efficiency sum lies below the stable-regime coefficient in 97.6% (SPF) and 97.0% (Livingston) of replications.

uncertainty reaches the results: if the regime parameters were poorly pinned down, the reclassified indicators would disagree across replications and the second-stage estimates would spread. Table 13 reports 999 replications, all successful. Appendix E.4 details the design.

The classification is stable: the median replication retains the baseline label for 93.1% of evaluation-window months. The professional volatile-regime sums are essentially unaffected, with intervals $[-0.897,\ -0.573]$ and $[-0.930,\ -0.524]$ strictly below zero. The stable-regime coefficients stay negative and moderate as the classification varies, and the volatile-regime sum lies below the stable coefficient in 97.6% of replications for the SPF and 97.0% for Livingston, so the gap between the two regimes survives the first-stage uncertainty. The adaptive loadings are just as stable. The volatile-regime household loading has a 95% interval of $[0.371,\ 0.432]$, against $[0.198,\ 0.254]$ for the SPF and $[0.118,\ 0.206]$ for Livingston, and the household loading exceeds the SPF's in 100.0% and Livingston's in 99.9% of replications. The volatile-regime loss differences remain negative in 97% of replications, with upper endpoints just above zero. Sampling uncertainty, not first-stage uncertainty, keeps the accuracy reversal from significance. The same reading applies to the rest of the table. Each interval measures how far the estimated classification can move an estimate, so an interval that excludes zero signals a stable estimate, not a significant one. The stable-regime loss differences illustrate the point: their intervals lie above zero, yet Table 3 shows them insignificant.

7 Conclusion

This paper examines whether the relationship between household and professional inflation expectations changes with the inflation environment. We compare median one-year-ahead expectations from the MSC, the SPF, and the Livingston Survey over 1981–2024, separating stable from volatile environments with a two-state Markov-switching model of realized CPI inflation.

The comparison has two dimensions, relative forecast accuracy and expectation formation, and they give different answers. On accuracy, neither group can be shown to outperform the other. On expectation formation, both the departures from rationality and the weight on recent inflation turn out to be state-dependent. Volatile-regime errors become predictable from the forecasts themselves, a pattern that is robust for the professionals and only suggestive for households. Households and professionals place remarkably similar weight on recently realized inflation when inflation is stable, and when inflation becomes volatile their responses diverge: households raise that weight while professionals lower it, leaving household expectations roughly 1.6 to 2 times as responsive to recent inflation as professional ones. The accuracy tests, the professional inefficiency, and the cumulative loadings hold up across survey frequencies, regime definitions, standard-error estimators, and the uncertainty of the estimated regime. We interpret these findings as consistent with state-dependent attention and extrapolation: as inflation turns volatile, the return to tracking it rises for households, while professionals may rely more heavily on models, policy anchors, and other forward-looking information.

The choice between the two surveys is therefore not a matter of accuracy. For central banks, the two series are better read together, with the gap between them as its own signal of how inflationary pressure is passing into expectations. In applied work, which measure enters a Phillips curve or a monetary-policy rule is a modeling assumption to state and justify.

Several limitations point to future work. Aggregate survey medians cannot speak to individual updating, disagreement, or compositional change. The estimated regime is a conditioning variable, not a treatment, so the results are descriptive rather than causal. The evidence, moreover, covers one country, one horizon, and few volatile-regime observations. Individual-level expectations, other countries and horizons, and direct measures of information acquisition could identify the mechanisms behind the divergence documented here.

References

Andrews DWK, Monahan JC (1992) An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. Econometrica 60:953-966. https://doi.org/10.2307/2951574

Ang A, Bekaert G, Wei M (2007) Do macro variables, asset markets, or surveys forecast inflation better? Journal of Monetary Economics 54:1163-1212. https://doi.org/10.1016/j.jmoneco.2006.04.006

Benchimol J, El-Shagi M, Saadon Y (2022) Do expert experience and characteristics affect inflation forecasts? Journal of Economic Behavior & Organization 201:205-226. https://doi.org/10.1016/j.jebo.2022.06.025

Bernanke BS (2007) Inflation expectations and inflation forecasting. Speech at the Monetary Economics Workshop of the NBER Summer Institute, Cambridge, MA, July 10.

Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31:307-327. https://doi.org/10.1016/0304-4076(86)90063-1

Bordalo P, Gennaioli N, Ma Y, Shleifer A (2020) Overreaction in macroeconomic expectations. American Economic Review 110:2748-2782. https://doi.org/10.1257/aer.20181219

Bracha A, Tang J (2025) Inflation levels and (in)attention. The Review of Economic Studies 92:1564-1594. https://doi.org/10.1093/restud/rdae063

Burke MA, Ozdagli A (2023) Household inflation expectations and consumer spending: evidence from panel data. The Review of Economics and Statistics 105:948-961. https://doi.org/10.1162/rest_a_01118

Candelon B, Roccazzella F (2025) Evaluating inflation forecasts in the euro area and the role of the ECB. Journal of Forecasting 44:978-1008. https://doi.org/10.1002/for.3235

Coibion O, Gorodnichenko Y (2015) Information rigidity and the expectations formation process: a simple framework and new facts. American Economic Review 105:2644-2678. https://doi.org/10.1257/aer.20110306

Coibion O, Gorodnichenko Y, Weber M (2022) Monetary policy communications and their effects on household inflation expectations. Journal of Political Economy 130:1537-1584. https://doi.org/10.1086/718982

D'Acunto F, Malmendier U, Ospina J, Weber M (2021) Exposure to grocery prices and inflation expectations. Journal of Political Economy 129:1615-1639. https://doi.org/10.1086/713192

D'Acunto F, Malmendier U, Weber M (2023) What do the data tell us about inflation expectations? In: Bachmann R, Topa G, van der Klaauw W (eds) Handbook of economic expectations. Elsevier, Amsterdam, pp 133-161. https://doi.org/10.1016/B978-0-12-822927-9.00012-4

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. Journal of Business & Economic Statistics 13:253-263. https://doi.org/10.1080/07350015.1995.10524599

El-Shagi M (2011) Inflation expectations: does the market beat econometric forecasts? The North American Journal of Economics and Finance 22:298-319. https://doi.org/10.1016/j.najef.2011.05.002

Evans M, Wachtel P (1993) Inflation regimes and the sources of inflation uncertainty. Journal of Money, Credit and Banking 25:475-511. https://doi.org/10.2307/2077719

Giacomini R, Rossi B (2010) Forecast comparisons in unstable environments. Journal of Applied Econometrics 25:595-620. https://doi.org/10.1002/jae.1177

Giacomini R, White H (2006) Tests of conditional predictive ability. Econometrica 74:1545-1578. https://doi.org/10.1111/j.1468-0262.2006.00718.x

Goodspeed T (2025) Trust the experts? The performance of inflation expectations, 1960–2023. International Journal of Forecasting 41:863-876. https://doi.org/10.1016/j.ijforecast.2024.06.006

Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57:357-384. https://doi.org/10.2307/1912559

Hansen BE (1992) Testing for parameter instability in linear models. Journal of Policy Modeling 14:517-533. https://doi.org/10.1016/0161-8938(92)90019-9

Harvey D, Leybourne S, Newbold P (1997) Testing the equality of prediction mean squared errors. International Journal of Forecasting 13:281-291. https://doi.org/10.1016/S0169-2070(96)00719-4

Henry E, Mulloy C, Sarte P-D (2023) What survey measures of inflation expectations tell us. Economic Brief 23-03, Federal Reserve Bank of Richmond.

Inclán C, Tiao GC (1994) Use of cumulative sums of squares for retrospective detection of changes of variance. Journal of the American Statistical Association 89:913-923. https://doi.org/10.1080/01621459.1994.10476824

Korenok O, Munro D, Chen J (2026) Inflation and attention thresholds. The Review of Economics and Statistics 108:842-850. https://doi.org/10.1162/rest_a_01402

Lamoureux CG, Lastrapes WD (1990) Persistence in variance, structural change, and the GARCH model. Journal of Business & Economic Statistics 8:225-234. https://doi.org/10.1080/07350015.1990.10509794

Lieb L, Schuffels J (2022) Inflation expectations and consumer spending: the role of household balance sheets. Empirical Economics 63:2479-2512. https://doi.org/10.1007/s00181-022-02222-8

Malmendier U, Nagel S (2016) Learning from inflation experiences. The Quarterly Journal of Economics 131:53-87. https://doi.org/10.1093/qje/qjv037

Mankiw NG, Reis R (2002) Sticky information versus sticky prices: a proposal to replace the New Keynesian Phillips curve. The Quarterly Journal of Economics 117:1295-1328. https://doi.org/10.1162/003355302320935034

Mankiw NG, Reis R, Wolfers J (2003) Disagreement about inflation expectations. NBER Macroeconomics Annual 18:209-248. https://doi.org/10.1086/ma.18.3585256

Mehra YP (2002) Survey measures of expected inflation: revisiting the issues of predictive content and rationality. Federal Reserve Bank of Richmond Economic Quarterly 88:17-36.

Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55:703-708. https://doi.org/10.2307/1913610

Newey WK, West KD (1994) Automatic lag selection in covariance matrix estimation. The Review of Economic Studies 61:631-653. https://doi.org/10.2307/2297912

Nyblom J (1989) Testing for the constancy of parameters over time. Journal of the American Statistical Association 84:223-230. https://doi.org/10.1080/01621459.1989.10478759

Pfajfar D, Santoro E (2010) Heterogeneity, learning and information stickiness in inflation expectations. Journal of Economic Behavior & Organization 75:426-444. https://doi.org/10.1016/j.jebo.2010.05.012

Reis R (2021) Losing the inflation anchor. Brookings Papers on Economic Activity 52:307-361. https://doi.org/10.1353/eca.2022.0004

Sansó A, Aragó V, Carrion-i-Silvestre JL (2004) Testing for changes in the unconditional variance of financial time series. Revista de Economía Financiera 4:32-53.

Sims CA (2003) Implications of rational inattention. Journal of Monetary Economics 50:665-690. https://doi.org/10.1016/S0304-3932(03)00029-1

Weber M, D'Acunto F, Gorodnichenko Y, Coibion O (2022) The subjective inflation expectations of households and firms: measurement, determinants, and implications. Journal of Economic Perspectives 36:157-184. https://doi.org/10.1257/jep.36.3.157

Weber M, Candia B, Afrouzi H, Ropele T, Lluberas R, Frache S, Meyer B, Kumar S, Gorodnichenko Y, Georgarakos D, Coibion O, Kenny G, Ponce J (2025) Tell me something I don't already know: learning in low- and high-inflation settings. Econometrica 93:229-264. https://doi.org/10.3982/ECTA22764

Appendices

A Additional Data Details

Table A.1 lists every series used in the paper, its source, and its treatment. Figure A.1 plots realized year-over-year CPI inflation over the sample period. The mean and, even more, the variability of the series shift over time, with turbulence concentrated in the early 1980s, around the 1990–1991 and 2008–2009 recessions, and during the post-pandemic surge. We fit the Markov-switching model of Section 2.2 to quarterly final-month year-over-year CPI inflation over 1981:Q2–2025:Q4. All series are final-vintage values as archived in the replication package. The Treasury-bill rate and the not-seasonally-adjusted CPI are not revised after release; the unemployment rate is subject to seasonal-factor revisions, so the origin dating of Section 2.3 concerns availability timing rather than literal real-time vintages.

B Regime Validation

The Markov-switching autoregression of Section 3 has a state-dependent innovation variance, so it partitions any sample into a high- and a low-variance state whether or not the variance of inflation genuinely shifts. The classification cannot validate itself. We therefore test the premise two ways. Table B.1 collects procedures that assume no Markov-switching data-generating process. Its first two panels never see the regime indicator. Its last two bring the indicator in as a regressor, Panel C to ask whether it absorbs the GARCH persistence and Panel D to ask whether it predicts the variance directly. Table B.2 asks whether the classification is an energy phenomenon.

We work with the residuals of an autoregression of year-over-year CPI inflation, with lag order chosen by the Bayesian information criterion (BIC): AR(6) for the 172 usable quarters of the classification window, and AR(13) for the 507 usable months of the monthly evaluation window. Filtering matters, because year-over-year inflation is strongly autocorrelated (at quarterly sampling the twelve-month windows overlap by nine months) and because the object of interest is the innovation variance, which is $\sigma^2_{s_t}$ in equation (2). We report results for demeaned but unfiltered inflation alongside.

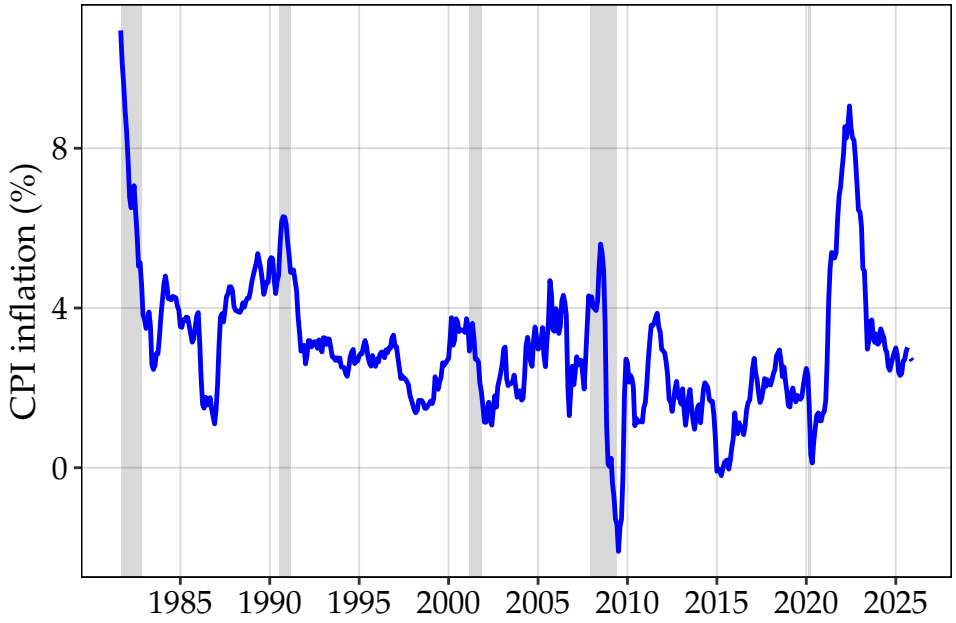
Which parameter is unstable? Panel A reports the Nyblom (1989) and Hansen (1992) L_c statistic, which tests constant parameters against martingale variation, with a sep-

Table A.1: Variables, Sources, and Transformations

Symbol	Series	Source	Frequency and treatment
π_t	CPI for All Urban Consumers, all items, not seasonally adjusted	FRED `CPIAUCNS`	Monthly, year-over-year percent change
π_t^{core}	CPI for All Urban Consumers, all items less food and energy, not seasonally adjusted	FRED `CPILFENS`	Quarterly, year-over-year percent change; core classification of Appendix B only
$\mathbb{E}_t \pi_{t+12}$ (MSC)	Median expected price change, next 12 months	University of Michigan, via FRED `MICH`	Monthly
$\mathbb{E}_t \pi_{t+12}$ (SPF)	Median one-year-ahead CPI inflation forecast	Federal Reserve Bank of Philadelphia, SPF `Inflation` file, variable `INFCPI1YR`	Quarterly, dated at the final month of the survey quarter
$\mathbb{E}_t \pi_{t+12}$ (Livingston)	Median forecast CPI levels	Federal Reserve Bank of Philadelphia, Livingston `medians` file, CPI sheet	Semiannual, annualized from forecast levels
i_t	Three-month Treasury bill, secondary market rate	FRED `TB3MS`	Monthly, percent per annum
$\mathcal{U}_t$	Civilian unemployment rate	FRED `UNRATE`	Monthly, percent, seasonally adjusted
Regime_t	Volatile-regime indicator	Authors' estimation of equation (2)	Quarterly, cutoff 0.5 on smoothed probabilities

Notes: The quarterly inflation series behind equation (2) and Appendix B is the year-over-year rate of each quarter's final month rather than a quarterly average, and the core series of Appendix B is built the same way; the 2025:Q4 observation therefore uses the December 2025 rate and is unaffected by the never-published October 2025 release. The Livingston expectation follows the Federal Reserve Bank of Philadelphia's documentation. Before June 1992, it is the annualized growth from the base-month CPI B, two months before the survey month, to the twelve-month-ahead forecast level F_{12}, $100[(F_{12}/B)^{12/14} - 1]$; from June 1992, when respondents begin estimating the survey-month CPI, it is the simple growth from that estimate to F_{12}. SPF and Livingston forecasts are dated at the final month of the survey period (March, June, September, or December, and June or December), so professional observation dates coincide with MSC months. Quarterly probabilities are dated at the quarter end, and each month within a quarter inherits the quarter's smoothed and filtered probabilities, and hence any classification derived from them.

Figure A.1: Year-over-Year U.S. CPI Inflation



Notes: Monthly year-over-year CPI inflation, 1981:M9–2025:M12. Shaded bars denote NBER recessions. The one-month gap in late 2025 is the never-published October 2025 CPI.

arate statistic for each parameter of the autoregression, the innovation variance included. None of the conditional-mean parameters, the intercept included, comes close to the 5% critical value of 0.470 at either frequency. The largest is 0.235. The variance statistic is 0.398 at quarterly frequency and 1.283 at monthly frequency against that same critical value. Over this sample, the conditional mean of inflation is stable and the conditional variance is not. The joint statistic rejects at neither frequency, as expected when one parameter of eight or fifteen is unstable: spreading the test across the stable coefficients dilutes its power, so the individual statistics are the informative object.

Because the score for the variance parameter, $e_t^2 - \sigma^2$, is not a martingale difference under conditional heteroskedasticity, we also report the variance statistic normalized by a Bartlett long-run variance. The monthly statistic survives the robustification at the 1% level, 0.905 against 0.748. The quarterly statistic does not, falling to 0.326. We return to that asymmetry below.

Is there a break in the unconditional variance? Panel B reports the cumulative-sum-of-squares statistic of Inclán and Tiao (1994) (IT) together with the two corrections of Sansó et al. (2004), written κ_1^S and κ_2^S to keep them distinct from the interest-rate coefficients of equation (9). The Inclán–Tiao statistic assumes independent innovations. The κ_1^S correction rescales it by the empirical fourth moment and so tolerates excess kurtosis. The κ_2^S correction divides by a Bartlett estimate of the long-run variance of e_t^2 and is therefore robust to GARCH. The gap between IT and κ_2^S is the diagnostic for the

objection that an apparent variance regime is only conditional heteroskedasticity. All three statistics have limiting distribution $\sup_r |W^*(r)|$, with quantiles 1.224, 1.358, and 1.628 from the closed-form Kolmogorov–Smirnov expression.

At monthly frequency, the evidence survives the correction with room to spare: on AR residuals the statistic falls from 3.751 to 2.002 under the GARCH-robust normalization and still exceeds the 1% critical value, with 1.945 under the data-driven bandwidth. At quarterly frequency, the evidence is only suggestive: $\kappa_2^S = 1.184$ at the baseline bandwidth, just below the 10% critical value, and 1.263 at the data-driven bandwidth, just above it. The Inclán–Tiao statistic, by contrast, is 2.487, significant at the 1% level. The correction for dependence in e_t^2 absorbs almost all of the quarterly evidence. Applied iteratively to the monthly residuals as the iterated-cumulative-sum-of-squares (ICSS) algorithm of Inclán and Tiao (1994), the κ_2^S test locates two change points, 2005:M3 and 2010:M12, with residual standard deviations of 0.221, 0.487, and 0.298 across the three implied segments. The quarterly procedure locates none.

Three things separate the two frequencies, and we do not read the quarterly non-rejection as overturning the monthly rejection. The quarterly sample is 172 observations against 507. The HAC correction is proportionately heavier there. Most importantly, the cumulative-sum-of-squares family is built against a single permanent break, whereas the classification has nine transitions. The partial-sum path returns toward zero each time the variance returns to its earlier level, which is precisely the configuration in which a supremum test loses power. At the frequency at which we estimate the Markov-switching model, the model-free evidence is therefore suggestive rather than established.

Do the located dates match the model? They do, without having been told to. The longest volatile episode in the classification runs from 2004:Q2 to 2011:Q2, twenty-nine quarters. The first monthly change point, 2005:M3, falls three quarters after the transition that opens it. The second, 2010:M12, falls three quarters before the transition that closes it. The two procedures share no parametric structure, one a filtered maximum-likelihood classification of an autoregression, the other a nonparametric sequential search on squared residuals.

Break or GARCH? Panel C runs the horse race directly, following Lamoureux and Lastrapes (1990). A Bollerslev (1986) GARCH(1,1) fitted to the monthly AR residuals returns $\hat{\alpha} + \hat{\beta} = 0.989$, the near-integrated persistence that neglected variance shifts are known to manufacture. Adding the ICSS segment dummies to the conditional-variance equation raises the log likelihood by 11.9 points, with $p < 0.001$ on two degrees of freedom, and cuts persistence to 0.630. Substituting the Markov-switching volatile dummy cuts it to 0.278. We do not lean on the second number: that dummy

comes from a model fitted to the same series under the same alternative, so its likelihood-ratio statistic is mechanically favored and is not a valid test. The ICSS row is the clean one and delivers the same answer more conservatively. The data contain discrete variance shifts, not GARCH persistence alone.

Does the classification predict variance? Panel D asks the complementary question, and it is the one that matters for the rest of the paper: does the regime label carry information about the size of inflation innovations? Regressing squared AR residuals on the volatile dummy gives $t = 4.95$ at monthly frequency and $t = 3.62$ at quarterly frequency, with residual standard deviations 1.82 and 2.37 times larger in the volatile state. The result survives restriction to the 2011-onward subsample, with $t = 2.14$ at monthly and $t = 2.53$ at quarterly frequency. That restriction matters, because the ICSS procedure does not isolate the post-2021 surge as a separate variance segment: the AR(13) filter absorbs much of 2021–2023 as a level movement, so the third ICSS segment spans 2011–2024 as a unit. Panel D, not Panel B, supports the classification of the post-2021 surge. Not every volatile episode is independently corroborated.

Is the volatile regime an energy phenomenon? The classification comes from headline inflation, and several of the episodes it selects are oil-price events. We therefore re-estimate equation (2) on core inflation, which excludes food and energy, over the same sample. Core inflation is quieter. Its stable-state innovation variance is 0.055 against 0.211 for headline inflation, its stable state is more persistent at $\widehat{p}_{11} = 0.991$, and it yields 24 volatile quarters against 64. The two classifications agree in 134 of the 178 quarters, and the core volatile quarters are nearly a subset of the headline ones, 22 of the 24 being volatile under both. Table B.2 reports the comparison.

Where the two part company is the informative part. Both call the early-1980s disinflation volatile, from 1981:Q3 to 1983:Q4, and both call the post-pandemic surge volatile, from 2020:Q4 to 2023:Q3. The 42 quarters that only headline inflation calls volatile are the 1986 oil-price collapse, the 1990 spike, the long commodity run-up through the financial crisis from 2004:Q2 to 2011:Q2, and 1984:Q1, the quarter in which the two classifications diverge at the end of the disinflation. The two quarters that only core calls volatile, 2020:Q2–2020:Q3, date the onset of the pandemic surge two quarters earlier than headline does. Energy therefore accounts for much of the volatile regime, but for neither of the two episodes that dominate the evaluation window. We retain the headline classification because the three surveys forecast headline inflation, so the regime belongs on the variable being forecast, but a reader who prefers a core-based reading should discount the middle of the sample rather than its ends.

What this appendix does and does not establish. It establishes three things. Over this sample, inflation has an unstable innovation variance and a stable conditional mean.

Table B.1: Variance Instability in Realized Inflation

Panel A: Nyblom–Hansen L_c parameter-constancy test

				Individual statistics		
Series	n	m	Joint L_c	$\max_j L_{\beta_j}$	L_{σ^2}	L_{σ^2}, HAC
Quarterly, AR(6)	172	8	1.172	0.230	0.398*	0.326
Monthly, AR(13)	507	15	3.143	0.235	1.283***	0.905***

Panel B: cumulative-sum-of-squares tests for a variance break

Series	IT	κ_1^S	κ_2^S	κ_2^S, NW bw	Break date
Quarterly, demeaned	2.771***	1.854***	1.024	1.179	2008:Q1
Quarterly, AR residuals	2.487***	1.307*	1.184	1.263*	2005:Q2
Monthly, demeaned	4.945***	3.508***	1.520**	1.256*	2008:M5
Monthly, AR residuals	3.751***	2.383***	2.002***	1.945***	2005:M3

Panel C: GARCH(1,1) persistence horse race, monthly AR(13) residuals

Conditional-variance specification	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}+\hat{\beta}$	LR	df	p
GARCH(1,1)	0.073	0.916	0.989	—	—	—
+ ICSS segment dummies	0.088	0.543	0.630	23.80	2	< 0.001
+ MS volatile dummy	0.169	0.109	0.278	25.24	1	< 0.001

Panel D: HAC regression of squared AR residuals on the volatile dummy

Sample	n	n volatile	$\hat{v}$	t	p	sd ratio
Monthly, full sample	507	177	0.111	4.95	< 0.001	1.82
Monthly, 2011 onward	168	42	0.078	2.14	0.033	1.37
Quarterly, full sample	172	58	0.959	3.62	< 0.001	2.37
Quarterly, 2011 onward	60	14	0.978	2.53	0.014	2.05

Notes: Year-over-year CPI inflation; the quarterly series is the classification window (1981:Q3–2025:Q4), the monthly series the evaluation window (1981:M9–2024:M12). $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Panel A. m counts parameters including σ^2; $\max_j L_{\beta_j}$ is the largest statistic across the intercept and the autoregressive coefficients. Hansen (1992) critical values: 0.353/0.470/0.748 for one parameter, 1.89/2.11/2.59 for $m = 8$, 3.26/3.54/4.07 for $m = 15$. The last column normalizes the variance score by a Bartlett long-run variance rather than by its outer product, at the Panel B baseline bandwidth.

Panel B. Bandwidths are $\lfloor 4(T/100)^{2/9} \rfloor$ (four quarterly, five monthly), and in the fourth column the automatic rule of Newey and West (1994); the break date is the argmax of the κ_2^S path. The iterative search applies κ_2^S recursively to the AR residuals by binary segmentation at the 5% level, with a minimum segment length of 24 months, and without the refinement pass of the original algorithm.

Panel C. Gaussian quasi-maximum likelihood with $h_t = \omega + z_t'\delta + \alpha e_{t-1}^2 + \beta h_{t-1}$ and h_1 set to the sample variance; log likelihoods -63.99, -52.09, and -51.37. ICSS dummies mark the change points 2005:M3 and 2010:M12; LR is the likelihood ratio against the first row. See the text on why the MS-dummy row is not a valid test.

Panel D. $\hat{v}$ is the coefficient on the volatile indicator, with Newey–West standard errors using VAR(1) prewhitening and the automatic bandwidth of Newey and West (1994); "sd ratio" compares residual standard deviations across regimes.

The instability is not an artifact of conditional heteroskedasticity at monthly frequency, and the quarterly evidence on that question is suggestive rather than established. Finally, a nonparametric search recovers both ends of the model's longest volatile episode. It does not establish that two states is the right number, that a first-order Markov chain is the right transition mechanism, that any particular quarter is correctly classified, or that the volatile episodes in the middle of the sample are anything other than energy-price events. Those remain modeling choices. Section 6.2 reports how the results change under alternative classifications built from the same model.

Table B.2: Headline and Core Classifications Compared

	Core stable	Core volatile	Total
Headline stable	112	2	114
Headline volatile	42	22	64
Total	154	24	178

Notes: Quarters of the 1981:Q3–2025:Q4 classification window, with the 1981:Q2 observation serving as the conditioning value. The headline classification is the one used throughout the paper. The core classification uses year-over-year inflation in the consumer price index for all items less food and energy, not seasonally adjusted, built to quarterly frequency the same way and estimated with the same specification, starting values, and software, with volatile quarters classified by the same 0.5 cutoff on smoothed probabilities. Its remaining estimates are a volatile-state innovation variance of 1.389, long-run means of 2.101 and 4.940, and a volatile-state transition probability of 0.901.

C Accuracy Test Specifications and Diagnostics

Section 4 concludes that neither survey can be shown more accurate. This appendix sets out the two tests behind that conclusion in full, then asks what happens below the 95% convention and which direction the data favor. All tests use the common-date loss differences of equation (4).

Equal accuracy conditional on origin-dated information. The conditional predictive ability test of Giacomini and White (2006) asks whether anything known at the origin predicts which survey will be more accurate. The test treats forecasts as primitives rather than as products of estimated models, the relevant case for survey expectations. The null is $\mathbb{E}[d_t \mid \mathcal{F}_t] = 0$. We regress d_t on a vector of instruments z_t and compute the Wald statistic that all coefficients are zero:

$$W = \hat{\beta}' \hat{V}^{-1} \hat{\beta}, \tag{C.1}$$

asymptotically χ^2 with $\dim(z_t)$ degrees of freedom, where $\hat{V}$ is a HAC estimate of the covariance of $\hat{\beta}$. We use two instrument sets, $\{1,\ \text{Regime}_t\}$ and $\{1,\ \text{Regime}_t,\ d_{t-s}\}$, with d_{t-s} the most recent loss difference verifiable at the origin.

The theory requires $z_t \in \mathcal{F}_t$, which is why Table 4 reports every test with the baseline indicator and with the filtered indicator $\text{Regime}_{f,t}$. Section 4.3 explains the distinction and reads the results. The note to Table 4 details the bootstrap complement, and Section 4.3 reports a third variant, the filtered indicator dated one quarter before the origin.

Equal accuracy at every point in the sample. A brief episode of significant dominance can still hide in a full-sample or full-regime mean. The fluctuation test of Giaco-

mini and Rossi (2010) checks for such an episode directly, computing the standardized rolling statistic:

$$\Phi_t \;=\; \hat{\sigma}^{-1} m^{-1/2} \sum_{j=t-m+1}^{t} d_j, \tag{C.2}$$

and comparing its entire path against critical values valid for the whole path under equal predictive ability. We use windows of $\mu = 0.3$ and 0.2 of each sample, $m = 52$ and 35 quarters for the MSC–SPF comparison and $m = 26$ and 17 half-years for MSC–Livingston, with $\hat{\sigma}^2$ the full-sample prewhitened Newey–West long-run variance at the one-year bandwidth. For the moving-block bootstrap bands, we demean the loss difference, resample it in one-year blocks, recompute the entire standardized path on each of 4,999 samples, and take the 95th percentile of the path maximum as the band, with the path-maximum p-value the exceedance share.

Figure C.1 plots all eight paths, each comparison under each loss at both window lengths. Under squared loss, the MSC–SPF path reaches 1.46 at $\mu = 0.3$ and 1.78 at $\mu = 0.2$, with a minimum of -1.02 in the shorter windows ending in 2024, against Giacomini–Rossi critical values of 3.01 and 3.18 and tighter moving-block bootstrap bands of 2.65 and 2.86. The bootstrap p-values for the path maximum are 0.651 and 0.562. The MSC–Livingston path peaks at 1.92 and 2.33, against bands of 2.94 and 3.22, with p-values of 0.456 and 0.354.[9]

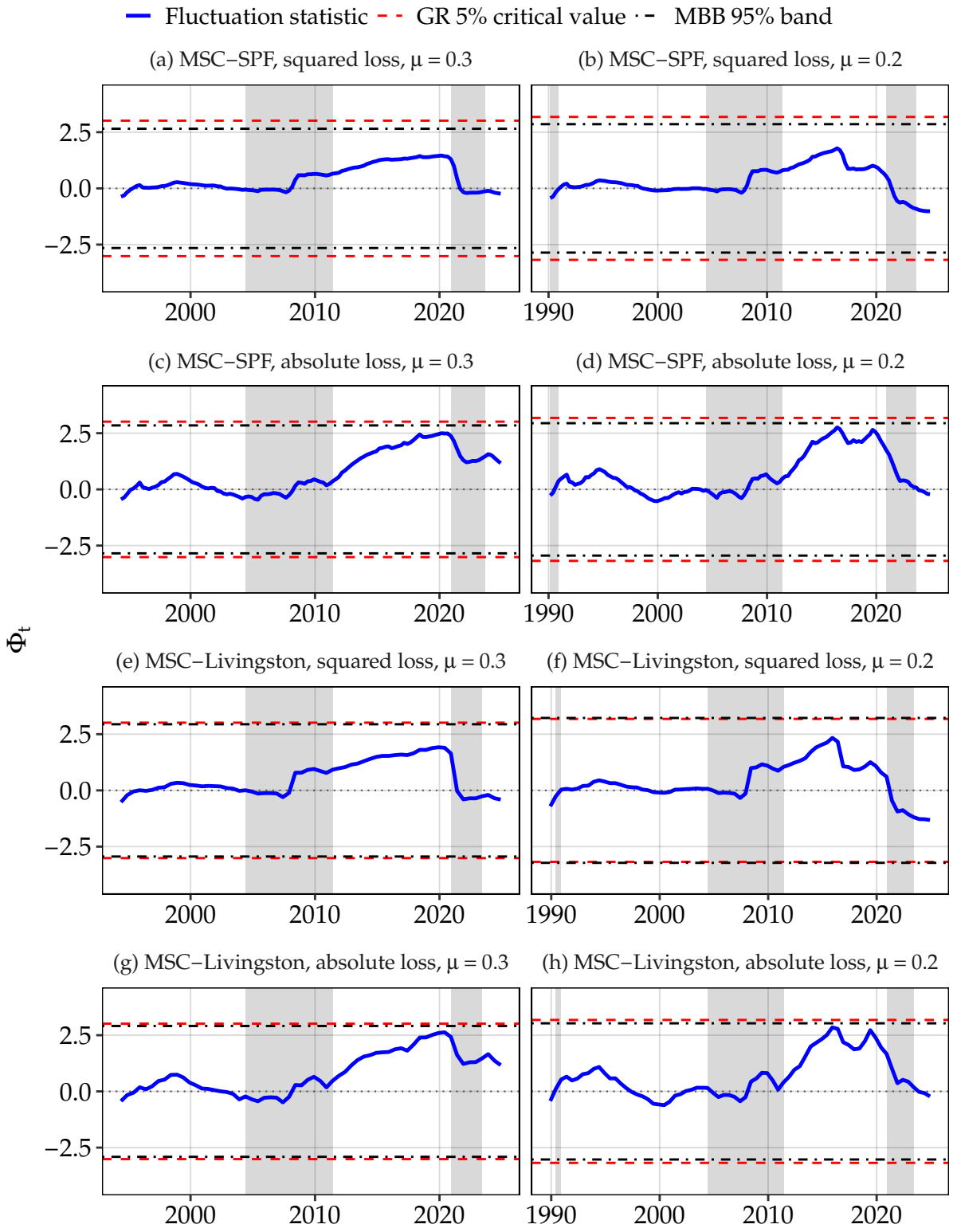
Absolute loss brings the paths closer to the bands without crossing them. The MSC–SPF maximum rises to 2.50 at $\mu = 0.3$ and 2.76 at $\mu = 0.2$, against bands of 2.84 and 2.94, and the MSC–Livingston maximum to 2.63 and 2.84, against bands of 2.91 and 3.03. The path-maximum bootstrap p-values are 0.127, 0.091, 0.102, and 0.080. All four peaks sit in windows that end between 2015 and 2020, dominated by the stable 2010s. This is the signal Table C.1 picks up below the 95% convention, a stable-regime professional advantage under absolute loss, and at its strongest it still stays inside every band.

Probing the answer. An answer as uniform as Section 4.2's invites two questions, one about the 95% convention and one about direction. Table C.1 collects the statistics behind both.

What happens below the 95% convention? At the 90% level one of the twelve intervals in Table C.1 moves off zero: the MSC–Livingston absolute-loss difference in the stable regime, whose 95% interval of $[-0.022,\ 0.356]$ in Table 3 tightens to $[0.005,\ 0.327]$. The interval moves off zero only at the shortest admissible block length. Blocks of three,

[9] The conclusion is also insensitive to the variance estimator: under the non-prewhitened Bartlett kernel used in the reference implementation of the test, the MSC–SPF path scales up, but at $\mu = 0.3$ its maximum, 2.25, remains inside the critical values.

Figure C.1: Giacomini–Rossi Fluctuation Tests



Notes: Each panel plots the rolling-window statistic Φ_t of equation (C.2), dated at the window end, at $\mu = 0.3$ and 0.2 of the common sample ($m = 52$ and 35 quarters for MSC–SPF, 26 and 17 half-years for MSC–Livingston); negative values favor households, positive values favor the professionals. Dashed horizontal lines are the two-sided 5% critical values of Giacomini and Rossi (2010); dash-dotted lines are moving-block bootstrap 95% bands for the path maximum (one-year blocks, 4,999 replications). Shaded areas mark periods classified as volatile by the baseline indicator.

Table C.1: Loss Differences: 90% Intervals and One-Sided Tail Areas

Comparison	Full sample	Stable regime	Volatile regime
MSC − SPF, squared loss	[−0.783, 0.982] 0.411 / 0.358	[−0.086, 0.907] 0.086 / 0.137	[−2.634, 1.794] 0.378 / 0.443
MSC − SPF, absolute loss	[−0.063, 0.272] 0.151 / 0.136	[−0.008, 0.335] 0.058 / 0.106	[−0.335, 0.321] 0.511 / 0.615
MSC − Livingston, squared loss	[−0.837, 0.976] 0.436 / 0.388	[−0.045, 0.790] 0.078 / 0.113	[−2.814, 1.948] 0.374 / 0.422
MSC − Livingston, absolute loss	[−0.072, 0.267] 0.165 / 0.147	[0.005, 0.327] [a] 0.044 / 0.083	[−0.373, 0.330] 0.473 / 0.544

Notes: Cells correspond one for one to the mean loss differences of Table 3. Brackets: the 90% moving-block bootstrap percentile interval, from the same 9,999 bootstrap draws as the 95% intervals of Table 3 (one-year blocks: four quarters for MSC–SPF, two half-years for MSC–Livingston). The two numbers beneath each interval are one-sided bootstrap tail areas: the share of the 9,999 bootstrap means that fall on the opposite side of zero from the point estimate, so a small value is directional support for the sign of the point estimate. The first tail area uses the one-year block length, the second a two-year block length (eight quarters / four half-years). [a]The only interval that excludes zero at the 90% level. It contains zero once the semiannual block length is raised to three, four, or six half-years ([−0.018, 0.349], [−0.035, 0.369], [−0.046, 0.387]).

four, or six half-years return it to containing zero at both levels. The Newey–West p-value is 0.26 to 0.28 at every bandwidth from two to six. That series also shows autocorrelation beyond the order the twelve-month overlap implies, with a data-driven bandwidth of 4.92 against the fixed two, which makes a longer block the better choice.

Which direction do the data favor? The one-sided bootstrap tail areas of Table C.1 speak to direction where the two-sided intervals cannot, and they are asymmetric across regimes. At the baseline block length, the tail area opposite the point estimate is 0.044 to 0.086 in all four stable-regime comparisons and 0.37 to 0.51 in all four volatile-regime comparisons. Lengthening the blocks to two years raises both, to 0.08–0.14 and 0.42–0.62. These tail areas measure directional support for the sign of the point estimate, not size under a null, so their levels are not p-values. The stable–volatile gap, a factor of three to eleven, persists at every block length we consider. What little directional support the data contain sits with the professionals in the stable regime.

D Additional Expectation-Formation Results

This appendix supports Section 5 with three exercises. Tables D.1 and D.2 remove the regime terms from every specification, to record what the familiar full-sample tests deliver and so to show which results exist only under regime conditioning. Table D.3 replaces the distributed lag with a single lag of inflation, to show that the loading divergence does not depend on the lag polynomial. Table D.4 estimates a revision-based alternative to the adaptive regression, to document the qualification reported in Section 5.2.

D.1 Results without regime interactions

Is anything in Section 5 visible without the regime split? Tables D.1 and D.2 answer by re-estimating the four rationality tests and the adaptive-expectations regression with no regime terms.

Without the interactions, the tests collapse into familiar full-sample results. No survey shows significant bias (Panel A). The own-forecast efficiency test rejects for both professional surveys but not for the MSC (Panel B). The bootstrap qualifies the panel twice: it does not confirm the SPF rejection and confirms the Livingston one at the 10% level only, and it places the MSC slope's interval strictly below zero even though the joint test does not reject on either route. No full-sample error persistence is significant (Panel C). The macroeconomic-information test does not reject for any survey (Panel D). What conditioning adds is the location of these failures: the full-sample efficiency rejections cannot be assigned to the volatile regime, and the household stable-regime departures, namely persistent errors and unexploited macroeconomic information, disappear entirely without it.

Unconditionally, the household cumulative loading on realized inflation, 0.404, is 1.8 times the SPF loading of 0.225 and 2.4 times the Livingston loading of 0.169. The macroeconomic controls stay jointly significant for all three surveys, so no survey follows a purely backward-looking rule.

D.2 Adaptive expectations with a single inflation lag

Does the loading divergence depend on the twelve-lag polynomial? Table D.3 answers by re-estimating the adaptive-expectations regression of Table 9 with a single lag of realized inflation, π_{t-1}, in place of the distributed lag $\beta(L)\pi_t$. The answer is no: the regime interaction remains positive and significant for the MSC and negative for both professional surveys, significantly so for the Livingston Survey.

D.3 An alternative test using forecast-error revisions

Does the regime divergence extend to the speed of error correction? Following Pfajfar and Santoro (2010), we answer with the alternative specification of Table D.4, in which forecast updating is captured by the latest verifiable forecast error:

$$\mathbb{E}_t[\pi_{t+12}] = \mathbb{E}_{t-s}[\pi_{t-s+12}] + (\zeta_1 + \zeta_2 \,\text{Regime}_t)\big(\pi_{t-s+12} - \mathbb{E}_{t-s}[\pi_{t-s+12}]\big) + \varepsilon_t, \quad \text{(D.1)}$$

Table D.1: Tests of Forecast Rationality without Regime Terms

	MSC	SPF	Livingston
Panel A: unbiasedness			
α	-0.307 (0.393) $[-0.674,\ 0.075]$	-0.088 (0.391) $[-0.445,\ 0.360]$	-0.065 (0.311) $[-0.446,\ 0.367]$
Panel B: own-forecast efficiency			
β	-0.482 (0.607) $[-0.909,\ -0.024]$	-0.536^{**} (0.250) $[-0.851,\ -0.181]$	-0.533^{***} (0.197) $[-0.866,\ -0.139]$
α	1.225 (1.537) $[-0.269,\ 2.661]$	1.497 (0.990) $[0.436,\ 2.564]$	1.491^{*} (0.783) $[0.370,\ 2.566]$
L6: Reject H_0: $\alpha = \beta = 0$?	F=0.320 p=0.727 p_{MBB}=0.897	F=3.498 p=0.032 p_{MBB}=0.104	F=5.324 p=0.007 p_{MBB}=0.058
Panel C: error persistence			
β	0.057 (0.198) $[-0.151,\ 0.267]$	0.183 (0.203) $[-0.015,\ 0.406]$	0.208 (0.157) $[-0.014,\ 0.465]$
α	-0.289 (0.396) $[-0.658,\ 0.096]$	0.005 (0.295) $[-0.327,\ 0.399]$	-0.045 (0.239) $[-0.406,\ 0.370]$
Panel D: use of macroeconomic information			
α	1.435 (1.329) $[-0.310,\ 3.161]$	1.367^{**} (0.638) $[-0.170,\ 2.845]$	1.221^{**} (0.569) $[-0.296,\ 2.712]$
β	-0.943 (0.634) $[-1.468,\ -0.403]$	-0.601^{*} (0.327) $[-1.412,\ 0.259]$	-0.665^{***} (0.244) $[-1.468,\ 0.166]$
γ	0.188 (0.160) $[-0.084,\ 0.474]$	0.114 (0.229) $[-0.176,\ 0.410]$	0.144 (0.178) $[-0.140,\ 0.441]$
κ	0.071 (0.190) $[-0.074,\ 0.210]$	-0.027 (0.163) $[-0.280,\ 0.211]$	-0.016 (0.143) $[-0.269,\ 0.225]$
δ	0.070 (0.105) $[-0.121,\ 0.277]$	0.014 (0.102) $[-0.207,\ 0.246]$	0.047 (0.097) $[-0.171,\ 0.276]$
L7: Reject H_0: $\gamma = \kappa = \delta = 0$?	F=1.062 p=0.365 p_{MBB}=0.359	F=0.107 p=0.956 p_{MBB}=0.962	F=0.350 p=0.789 p_{MBB}=0.879

Notes: Each panel estimates the corresponding specification of Section 2.3, equations (6)–(9), with all regime terms (the regime dummy and its interactions) removed. Samples: MSC 519 monthly (1981:M9–2024:M12), SPF 174 quarterly, Livingston 87 semiannual. Prewhitened Newey–West standard errors in parentheses (bandwidth of one year: 12, 4, and 2 lags). Panel C uses 170 SPF observations (the lagged own error is unavailable at the start of the SPF sample). The October 2024 MSC forecast error is unavailable because the October 2025 CPI was never published. Brackets: residual moving-block bootstrap 95% percentile intervals from 4,999 replications; p_{MBB} beneath a joint test: its null-imposed residual-MBB p-value from 1,999 replications. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table D.2: Adaptive Expectations without Regime Terms

	MSC	SPF	Livingston
$\beta(1)$	0.404^{***}	0.225^{***}	0.169^{***}
	(0.023)	(0.038)	(0.042)
	[0.345, 0.459]	[0.149, 0.302]	[0.084, 0.254]
α	1.880^{***}	0.471	0.511
	(0.146)	(0.382)	(0.348)
	[1.599, 2.174]	[0.076, 0.859]	[0.070, 0.929]
λ_1	-0.071^{***}	0.196	0.177
	(0.020)	(0.171)	(0.180)
	[−0.136, 0.001]	[−0.087, 0.473]	[−0.218, 0.568]
λ_2	0.116^{***}	−0.047	−0.018
	(0.022)	(0.138)	(0.148)
	[0.042, 0.182]	[−0.298, 0.216]	[−0.384, 0.348]
θ_1	−0.025	0.330^{*}	0.580^{**}
	(0.118)	(0.170)	(0.229)
	[−0.190, 0.138]	[0.047, 0.596]	[0.092, 1.070]
θ_2	−0.029	−0.084	−0.321
	(0.111)	(0.156)	(0.224)
	[−0.192, 0.131]	[−0.351, 0.199]	[−0.806, 0.161]
L8: Reject H_0: $\lambda_1 = \lambda_2 = \theta_1 = \theta_2 = 0$?	F=11.427	F=39.630	F=25.194
	p<0.001	p<0.001	p<0.001
	p_{MBB}=0.011	p_{MBB}=0.001	p_{MBB}=0.001
R^2	0.672	0.886	0.860
Adjusted R^2	0.661	0.875	0.829

Notes: Estimates of equation (10) with the regime interaction removed. Samples: MSC 520 monthly (1981:M9–2024:M12), SPF 174 quarterly, Livingston 87 semiannual. The reported $\beta(1)$ is the cumulative loading. Prewhitened Newey–West standard errors in parentheses (bandwidth of one year: 12, 4, and 2 lags). Brackets: residual moving-block bootstrap 95% percentile intervals from 4,999 replications; p_{MBB} beneath a joint test: its null-imposed residual-MBB p-value from 1,999 replications. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table D.3: Adaptive Expectations with a Single Inflation Lag

	MSC	SPF	Livingston
	0.263^{***}	0.250^{***}	0.257^{***}
β_1	(0.038)	(0.073)	(0.057)
	[0.196, 0.328]	[0.166, 0.335]	[0.154, 0.359]
	0.093^{***}	-0.030	-0.070^{**}
β_2	(0.027)	(0.046)	(0.034)
	[0.049, 0.138]	[−0.088, 0.028]	[−0.138, −0.002]
	0.356^{***}	0.220^{***}	0.186^{***}
$\beta_1 + \beta_2$	(0.024)	(0.041)	(0.039)
	[0.310, 0.401]	[0.161, 0.281]	[0.121, 0.255]
	2.085^{***}	0.380	0.354
α	(0.136)	(0.443)	(0.389)
	[1.795, 2.373]	[−0.028, 0.784]	[−0.109, 0.798]
	-0.069^{***}	0.215	0.156
λ_1	(0.017)	(0.152)	(0.148)
	[−0.137, 0.006]	[−0.062, 0.482]	[−0.213, 0.509]
	0.112^{***}	-0.055	0.012
λ_2	(0.022)	(0.125)	(0.130)
	[0.037, 0.181]	[−0.301, 0.198]	[−0.322, 0.356]
	-0.069	0.322^{*}	0.735^{***}
θ_1	(0.115)	(0.163)	(0.225)
	[−0.225, 0.089]	[0.056, 0.574]	[0.258, 1.227]
	0.046	-0.076	-0.495^{**}
θ_2	(0.111)	(0.155)	(0.220)
	[−0.110, 0.200]	[−0.327, 0.188]	[−0.985, −0.016]
R^2	0.675	0.884	0.856
Adjusted R^2	0.671	0.880	0.846

Notes: Estimates of equation (10) with both lag polynomials collapsed to scalars, $\mathbb{E}_t[\pi_{t+h}] = \alpha + (\beta_1 + \beta_2 \,\text{Regime}_t)\,\pi_{t-1} + \text{controls}$, where π_{t-1} is the one-month lag for all three surveys and $\beta_1 + \beta_2$ is the volatile-regime loading. Samples: MSC 520 monthly (1981:M9–2024:M12), SPF 174 quarterly, Livingston 87 semiannual. Prewhitened Newey–West standard errors in parentheses (bandwidth of one year: 12, 4, and 2 lags). Brackets: residual moving-block bootstrap 95% percentile intervals from 4,999 replications. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

where ζ_1 is the baseline updating speed in the stable regime and ζ_2 the change in the volatile regime. We impose the unit coefficient on the earlier forecast, so we regress the forecast revision on the interacted verifiable error, without an intercept. The revision lag s is the smallest multiple of the survey's frequency such that the earlier forecast's target period precedes t, so that the forecast is verifiable when the time-t forecast is formed. The lag is thirteen months for the monthly MSC (so the error term is $\pi_{t-1} - \mathbb{E}_{t-13}\pi_{t-1}$), fifteen months for the quarterly SPF, and eighteen months for the semiannual Livingston Survey.

We detect no such extension. In the stable regime, the updating speed ζ_1 is positive and significant for the MSC and the SPF, and positive but insignificant for the Livingston Survey. In the volatile regime, the sums $\zeta_1 + \zeta_2$ are positive and significant for all three surveys. The household speed rises from 0.261 to 0.398, the SPF's is essentially unchanged, and the Livingston speed rises from 0.125 to 0.241. The household rise points the same way as Table 9, where households weight recent inflation more heavily once it turns volatile, but the professional decline of Table 9 has no counterpart here, and the Livingston rise even points the other way. The shifts ζ_2 are statistically insignificant for every survey, so the revision test cannot establish any regime dependence. The two findings do not contradict each other, because the two regressions measure different objects. The adaptive regression asks how the forecast level weights recently realized inflation. The revision test asks how fast the forecast corrects its latest verifiable error. A forecaster can lower the weight on recent inflation when it turns volatile and still correct verifiable errors at an unchanged speed. The revision test is also coarser than the adaptive regression, since it condenses updating into a single verifiable error in place of the distributed lag.

E Estimator Diagnostics and Bootstrap Procedures

This appendix supports Sections 6.3 and 6.4.

E.1 Serial correlation in the loss difference

The twelve-month horizon makes d_t a moving average of order $q = h/\text{freq} - 1$ even when both forecasts are optimal: $q = 3$ for the quarterly MSC–SPF comparison and $q = 1$ for the semiannual MSC–Livingston comparison. Autocorrelation up to lag q is mechanical. What matters for inference is autocorrelation beyond lag q, which the overlap does not explain. Table E.1 reports the diagnostic.

Panel A shows that the serial correlation present is very nearly what the overlap

Table D.4: Forecast-Error-Revision Test of Adaptive Expectations

	MSC	SPF	Livingston
ζ_1	0.261^{***}	0.240^{***}	0.125
	(0.061)	(0.073)	(0.077)
	[0.126, 0.434]	[0.144, 0.408]	[−0.071, 0.473]
ζ_2	0.138	0.005	0.116
	(0.088)	(0.085)	(0.131)
	[−0.064, 0.304]	[−0.179, 0.111]	[−0.272, 0.328]
$\zeta_1 + \zeta_2$	0.398^{***}	0.245^{***}	0.241^{**}
	(0.065)	(0.049)	(0.103)
	[0.290, 0.498]	[0.177, 0.304]	[0.072, 0.359]
R^2	0.347	0.436	0.172
Adjusted R^2	0.345	0.429	0.153

Notes: Estimates of equation (D.1), without an intercept; R^2 is uncentered. Samples: MSC 520 monthly (1981:M9–2024:M12), SPF 169 quarterly, Livingston 87 semiannual. The MSC and Livingston samples are unchanged because both surveys supply pre-window forecasts for the lagged term; the SPF's CPI question begins in 1981:Q3, so its fifteen-month lag costs the first five quarters. Prewhitened Newey–West standard errors in parentheses (bandwidth of one year: 12, 4, and 2 lags). Brackets: residual moving-block bootstrap 95% percentile intervals from 4,999 replications. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

predicts, and Section 6.3 reads the pattern off it: in both series the excess-autocorrelation test does not reject. The variance inflation reported there is substantial for the quarterly comparison, a factor of 6.7 under squared loss over the full sample.

Panel B converts that variance inflation into the quantity that matters for power. The effective sample sizes are small, most severely for the MSC–SPF squared-loss difference in the volatile regime, where 64 nominal observations carry the information of about ten independent ones.

E.2 Prewhitening and small-sample corrections

Prewhitening. The Andrews–Monahan prefilter is worth its variance cost only when the score series is persistent. Our score series is persistent. The first-order autocorrelation of the loss difference is 0.742 for the quarterly comparison, with a t statistic of 14.6, and 0.300 for the semiannual one, with a t statistic of 3.0. The quarterly series is squarely in the case prewhitening was designed for. The semiannual series is persistent but only moderately so, and there prewhitening moves the standard error only modestly. In both series, prewhitening raises the standard error relative to the textbook estimator, so it is the conservative choice.

Small-sample corrections. Three conventions are at issue. The baseline draws p-values from a t distribution with residual degrees of freedom rather than from the normal, and it reports moving-block bootstrap intervals alongside every test, which is itself a small-sample device that does not rely on asymptotic approximation. The

Table E.1: Serial Correlation in the Loss Difference d_t

Panel A: autocorrelation structure, full sample

			Autocorrelation			p-value			
Result	n	q	ρ_1	ρ_2	ρ_3	LB(8)	Excess	VIF	$\widehat{bw}$
MSC−SPF loss difference	174	3	0.74	0.41	0.17	< 0.001	0.442	6.74	7.67
MSC−Livingston loss difference	87	1	0.30	0.02	0.05	0.351	0.230	1.64	2.96

Panel B: effective sample size by regime

	Stable regime			Volatile regime		
Result	n	VIF	eff. n	n	VIF	eff. n
MSC−SPF loss difference	110	4.80	23	64	6.47	10
MSC−Livingston loss difference	55	2.08	27	32	1.43	22

Notes: Squared-loss differences, matching the scope of Section 6. $q = h/\text{freq} - 1$ is the moving-average order implied by the twelve-month overlap alone. ρ_j is the sample autocorrelation of d_t at lag j; the 5% Bartlett band is ± 0.149 for $n = 174$ and ± 0.210 for $n = 87$. LB(8) is the Ljung–Box p-value at eight lags, computed on the residuals from the regime-mean regression, the series that enters the inference. "Excess" is the p-value of a joint test that the coefficients on lags $q+1$ through $q+4$ are zero in a regression of d_t on an intercept and those four lags only; the mechanical MA(q) correlation at lags $\leq q$ remains in the error and is absorbed by the prewhitened Newey–West covariance at the baseline bandwidth. VIF is the variance inflation factor, the ratio of the prewhitened Newey–West variance of the mean to its i.i.d. counterpart; effective n is the nominal sample size divided by the unrounded VIF and then rounded. In Panel B, the numerator is the regime-mean coefficient variance from the full-sample regime regression that underlies Table 3, at the baseline bandwidth, and the denominator is the regime's own variance divided by its n; the regime subsamples are not extracted and recomputed in isolation. $\widehat{bw}$ is the data-driven bandwidth of Newey and West (1994), for comparison with the fixed one-year bandwidths of four and two used in the baseline.

baseline does not apply the $n/(n-k)$ scaling of the covariance estimator, and it does not apply the Harvey et al. (1997) correction to the Diebold–Mariano statistic:

$$DM^* = DM \times \sqrt{\frac{n+1-2h+h(h-1)/n}{n}} \sim t_{n-1}, \tag{E.1}$$

where h is the horizon in sampling units ($h = 4$ quarters, $h = 2$ half-years). Table E.2 adds both the $n/(n-k)$ scaling and the Harvey–Leybourne–Newbold correction. Neither changes any conclusion. The $n/(n-k)$ scaling moves p-values by at most 0.006. The Harvey–Leybourne–Newbold correction combines shrinkage factors of 0.980 and 0.983 with a truncated variance estimator smaller than the prewhitened Newey–West one, and it lowers the p-value in four of the six cells and leaves one unchanged. In the remaining cell, the MSC–Livingston volatile-regime difference, the shrinkage factor dominates the slightly smaller variance and the p-value rises by 0.018. Wherever the correction moves a p-value materially, it moves it down, so the baseline remains the more conservative convention. Even so, the smallest p-value anywhere in the table is 0.224. No cell approaches significance under any combination. The gaps that make the regime cells non-contiguous break the moving-average structure, so the correction is an approximation there: we compute the truncated autocovariances and n on the regime's observations placed side by side, with h unchanged, which if anything overstates the dependence.

Table E.2: Estimator and Small-Sample Sensitivity of the Loss-Difference Tests

Result	Sample	$\bar{d}$	Baseline p	p, adjust $= T$	p, HLN corrected
MSC−SPF loss difference	Full sample	0.038	0.965	0.965	0.953
	Stable regime	0.405	0.324	0.327	0.264
	Volatile regime	−0.592	0.787	0.788	0.714
MSC−Livingston loss difference	Full sample	0.009	0.988	0.988	0.988
	Stable regime	0.357	0.270	0.276	0.224
	Volatile regime	−0.588	0.704	0.707	0.722

Notes: Squared-loss differences, matching the scope of Section 6; $\bar{d}$ reproduces Table 3. "Baseline" is the prewhitened Newey–West estimator at the fixed one-year bandwidth with `adjust = FALSE`, referred to t_{n-k}, as used throughout the paper. "adjust $= T$" applies the $n/(n-k)$ finite-sample scaling to the same estimator. "HLN corrected" applies equation (E.1) to a Diebold–Mariano statistic built from the truncated variance estimator $\hat{\gamma}_0 + 2\sum_{k=1}^{h-1} \hat{\gamma}_k$ of Harvey et al. (1997), referred to t_{n-1}.

E.3 Residual moving-block bootstrap for regression quantities

Because one-year-ahead forecast errors overlap, HAC inference may be inaccurate in samples of this size. For every regression quantity reported in the paper's coefficient tables, we therefore run a residual moving-block bootstrap. Two independent implementations in the replication package produce the eleven quantities of Table E.3 and agree digit for digit. Holding the regressors fixed, we resample residual blocks of one year (twelve monthly, four quarterly, or two semiannual observations) with replacement, rebuild the dependent variable from the fitted values plus the resampled residuals, and re-estimate the model on each of 4,999 samples. We report 95% percentile intervals. For each labeled joint test (L1–L8), the accompanying p_{MBB} is instead a bootstrap p-value from 1,999 replications. We draw residual blocks from the null-imposed model, the restricted least-squares fit under that test's constraints, and recompute the prewhitened Newey–West F statistic in each replication. Block start points are drawn uniformly with replacement, the concatenated series is truncated to the sample length, and bootstrap p-values use the convention $(1 + \#\{F_b \geq F_0\})/(1 + B)$, with the rare replication whose restricted re-estimation fails discarded.

Table E.3 reports the results. The bootstrap confirms both halves of the professional efficiency contrast and the formation results: the volatile-regime inefficiency intervals, the six cumulative-loading intervals, and the household loading-shift interval all exclude zero, while the stable-regime efficiency intervals straddle it.

We also use the moving-block bootstrap, with the same one-year block lengths, for the pairwise loss-difference intervals of Table 3, with 9,999 replications, resampling the loss difference and the regime indicator jointly in blocks, which preserves the dependence between relative performance and the regime.

Table E.3: Moving-Block Bootstrap Inference for Key Coefficients

Result	Newey–West estimate	MBB 95% interval
SPF own-forecast efficiency, stable (β_1)	-0.230 (s.e. 0.272)	$[-0.681,\ 0.274]$
Livingston own-forecast efficiency, stable (β_1)	-0.173 (s.e. 0.222)	$[-0.641,\ 0.335]$
SPF own-forecast efficiency, volatile ($\beta_1 + \beta_2$)	-0.831 (s.e. 0.288)	$[-1.223,\ -0.415]$
Livingston own-forecast efficiency, volatile ($\beta_1 + \beta_2$)	-0.871 (s.e. 0.245)	$[-1.285,\ -0.399]$
MSC adaptive loading, stable ($\beta_1(1)$)	0.305 (s.e. 0.038)	$[0.227,\ 0.382]$
SPF adaptive loading, stable ($\beta_1(1)$)	0.284 (s.e. 0.070)	$[0.185,\ 0.388]$
Livingston adaptive loading, stable ($\beta_1(1)$)	0.302 (s.e. 0.060)	$[0.185,\ 0.410]$
MSC adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.383 (s.e. 0.019)	$[0.329,\ 0.437]$
SPF adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.242 (s.e. 0.040)	$[0.168,\ 0.317]$
Livingston adaptive loading, volatile ($\beta_1(1) + \beta_2(1)$)	0.193 (s.e. 0.044)	$[0.112,\ 0.271]$
MSC adaptive loading shift ($\beta_2(1)$)	0.079 (s.e. 0.030)	$[0.033,\ 0.125]$

Notes: Newey–West estimates are taken from Tables 6 and 9. The bootstrap procedure is the one specified in the text of this appendix.

E.4 Parametric bootstrap for regime-estimation uncertainty

We estimate the regime indicator Regime_t in a first-stage maximum-likelihood fit of the Markov-switching model (2), so second-stage standard errors understate total uncertainty. The following parametric bootstrap quantifies the first-stage contribution. We apply it to a small set of core quantities rather than to every coefficient:

1. Simulate a pseudo quarterly inflation path of the classification window's length from the estimated MS-AR(1), using the point estimates of $(\alpha_s, \rho_s, \sigma_s^2)$ and the estimated transition matrix, after a 100-quarter burn-in.

2. Re-estimate the model on the simulated path from two starting values, the values used for the actual-data estimation and the actual-data point estimates. Retain the fit with the higher likelihood and label its higher-variance state volatile.

3. Using the re-estimated parameters, recompute smoothed probabilities on the actual inflation series and reclassify the sample at the baseline cutoff of 0.5, rebuilding the monthly indicator by the same quarter-end inheritance rule as the baseline.

4. Re-run the second stage on the actual data with the reclassified regime indicator, and store the core quantities.

The core quantities are the four squared-loss differences by regime, the professional own-forecast coefficients in the stable regime and their volatile-regime sums, and the stable and volatile cumulative loadings of the three surveys. Because we hold the data fixed and only the classification varies, the resulting intervals isolate first-stage uncertainty and complement, rather than replace, the sampling-based inference of the main text.

Table 13 in Section 6.4 reports the results from 999 replications. Estimation succeeded in all 999 replications, so we never used the rule that discards a replication whose estimation fails at both starting values or whose smoothing of the actual series fails. A quantity can still be undefined within a successful replication. When the reclassified indicator takes a single value over a survey's estimation sample, or leaves one regime too few dates to identify the regime interactions, the interaction coefficients drop out; we then treat the survey's regime-specific coefficients as undefined rather than keep a stable-regime coefficient that the volatile dates helped determine. The loss difference of a regime that never occurs likewise has no dates to draw on. That leaves each entry of Table 13 defined in between 978 and 999 replications: the volatile-regime loss differences are defined in all 999, and the smallest count, 978, belongs to the two Livingston loading rows.