Adapting to Inflation Uncertainty

Luke McDougall

Working Paper No. 006 March 2026

EMAP Working Paper Series

Queen Mary University of London

Adapting to Inflation Uncertainty

Forecasting UK CPI Inflation with Unobserved Components Models: A Stochastic Volatility Approach

Non-technical summary

UK inflation has been highly volatile in recent years, surging to double digits in 2022 before falling back but remaining above target. These swings exposed weaknesses in forecasting models, which often failed to anticipate both the speed of the rise and the pace of the subsequent decline. The Bank of England and other policymakers have recognised the need for forecasting tools that are more flexible and better capture the uncertainty around future outcomes.

This paper tests whether an advanced statistical model (UCSV) can improve short-term inflation forecasts compared with standard methods (AR benchmark). The study examines UK consumer price inflation from 1996 to 2025, covering overall inflation and its 12 components (e.g., food, transport, housing). Forecasts are tested across three periods: the stable pre-COVID years (2016–2019), the pandemic and energy shock (2020–2023), and the recent post-COVID period (2024–2025). Two approaches are compared: forecasting overall inflation directly, versus forecasting individual components and then combining them.

The advanced model consistently outperforms standard methods beyond one month, delivering smaller forecast errors and more realistic uncertainty estimates. The gains are especially strong when forecasting overall inflation directly, which proves robust across all three periods. Forecasting individual components works well at short horizons, but performs worse at longer horizons where combining volatile components amplifies noise rather than improving accuracy.

These findings have practical implications. Forecast errors remain large at the one-month horizon, highlighting that very short-term prediction is inherently difficult and expert judgement remains essential. The results suggest that while no single model dominates, this approach provides a useful addition to the forecasting toolkit, particularly valuable for communicating uncertainty during volatile periods.

Introduction & Motivation

UK CPI inflation rose above the Bank of England's 2% target in mid-2021 and reached 11.1% in October 2022. The surge reflected the reopening from the pandemic, global supply bottlenecks, and the energy and food price shock following Russia's invasion of Ukraine. Forecasts during this episode repeatedly misjudged both the ascent and the subsequent disinflation, prompting an independent review of the Bank of England's forecasting framework by Dr Bernanke (2024). His review recommended modernising the forecasting toolkit, strengthening short-term models, improving systematic real-time evaluation and placing greater emphasis on communicating uncertainty. This study speaks directly to those recommendations by assessing whether a state-of-the-art unobserved-components model with stochastic volatility (UCSV) can deliver more robust short-horizon forecasts for UK inflation and better capture time-varying uncertainty.

Although inflation has fallen back from its peak, risks remain material. UK headline CPI was the highest in the G7 for the fifth consecutive month, standing at 3.8% in July 2025 compared with 2.0% in the euro area and 2.7% in the US. A key source of this divergence is services inflation, which remains elevated in the UK at 5.0% against 3.6% in the US and 3.2% in the euro area, with higher energy and food inflation also contributing. According to the August 2025 Monetary Policy Report,

inflation is projected at around 4% in September and is not expected to return to target until later in the forecast horizon. This reinforces the urgency of using forecasting methods that can both detect turning points and quantify uncertainty in a timely way, while remaining robust out-of-sample. The recent episode highlighted not only large point-forecast errors but also density forecasts that understated the probability of extreme outcomes, underscoring the importance of evaluating models on their ability to capture changing uncertainty.

The academic literature makes clear that inflation is notoriously difficult to forecast out-of-sample, with traditional benchmarks often proving surprisingly hard to beat. Atkeson and Ohanian (2001) show that a very simple approach, using the latest 12-month inflation rate as the forecast for the next year, competes closely with Phillips-curve specifications at short horizons. Stock and Watson (2007, 2010, 2019) demonstrate that models which allow for a drifting mean and time-varying volatility often match or outperform more elaborate alternatives. The UCSV model builds directly on these insights, offering a framework that captures persistent shifts in the trend, evolving seasonal patterns, and changing volatility - features that became especially salient during the pandemic and energy-price shocks.

The objective of this study is to provide a systematic assessment of short-horizon inflation forecasts using the UCSV model, benchmarked against a standard autoregressive approach. Forecasts are generated in a recursive, pseudo-real-time setting with an expanding window and evaluated at horizons of one to six months ahead. To align with policy practice, all results are reported in year-on-year CPI inflation. Performance is judged on both point forecast accuracy, using RMSEs and Diebold–Mariano tests, and density performance, using log predictive scores and other metrics. The analysis is carried out across three distinct regimes: a pre-COVID stable period (2016–2019), the COVID and energy-shock period (2020–2023), and the recent post-COVID/energy hock period (2024–2025), capturing the contrasting challenges of forecasting in each environment.

Literature review

Inflation forecasting is notoriously difficult and a substantial literature finds that simple specifications often provide strong out-of-sample performance. Atkeson and Ohanian (2001) showed that a random-walk model, using the latest 12-month inflation rate as the forecast for the next year, performs as well as or better than Phillips-curve forecasts at horizons up to a year in the United States. Stock and Watson (2007) argued that inflation predictability is dominated by a slowly varying local mean and short-lived noise, which limits the gains from adding predictors. Their subsequent work emphasised that univariate benchmarks are surprisingly difficult to beat, a conclusion echoed in the Bernanke (2024) review, which noted that autoregressive or random-walk type models often perform as well as more sophisticated methodologies.

Unobserved-components models provide a natural way to capture the features highlighted by these studies. The UCSV specification decomposes inflation into a stochastic trend and a transitory component, with the variances of trend and irregular innovations following random walks in logs. This structure allows for gradual shifts in the level of inflation and time-varying uncertainty, features that became salient in the past decade. Stock and Watson (2010, 2016) formalised and applied the UCSV model to US and euro area inflation, showing robust performance across subsamples and measures of inflation. The key contribution of these models is their ability to combine flexibility in the trend with evolving volatility and seasonal dynamics, while remaining parsimonious and relatively easy to update and maintain - qualities that are particularly valued in policy settings.

Forecast evaluation has traditionally focused on point accuracy, with root mean squared error (RMSE) and statistical tests such as the Diebold–Mariano test (Diebold & Mariano, 1995) forming the standard tools. More recent work stresses that density forecasts are equally important, since policymakers care about the entire distribution of possible outcomes, not just the mean. Standard approaches include the

log predictive score (Amisano & Giacomini, 2007), the continuous ranked probability score (CRPS; Gneiting & Raftery, 2007), and calibration checks such as the probability integral transform (Diebold, Gunther & Tay, 1998). The emphasis on density forecasting also aligns with Bernanke (2024), who argued that better communication of uncertainty is a key part of improving the credibility of forecasts.

An additional question concerns the level of aggregation. A long-standing issue is whether it is better to forecast inflation aggregates directly, or to forecast disaggregated inflation rates and then re-aggregate. Theoretically, if the data-generating process is known, aggregating disaggregate forecasts must be at least as good as forecasting the aggregate directly (Lütkepohl, 1987). In practice, however, once the process has to be estimated, the relative performance of the two approaches becomes uncertain. A direct aggregate forecast may be more accurate, since it requires estimating fewer parameters. If the disaggregates share similar dynamics, forecasting aggregates directly is likely to work better in small samples; if the disaggregates are highly persistent, then bottom-up aggregation may perform better. Ultimately, this is an empirical question. For the euro area, Hubrich (2005) finds that neither approach necessarily dominates, while Bermingham and D'Agostino (2011) report results more favourable to bottom-up aggregation. In line with this literature, this study compares top-down and bottom-up UCSV forecasts for UK CPI.

Alternative approaches include Bayesian vector autoregressions (BVARs) and machine learning methods. BVARs form a key component of the Bank of England's forecasting toolkit, with shrinkage priors enhancing their robustness. Domit, Monti, and Sokol (2016) demonstrate that a BVAR benchmark performs comparably to the Bank's DSGE model COMPASS in forecasting inflation. More recently, Brignone and Piffer (2025) propose a structural BVAR framework that effectively models UK and global shocks, offering improved forecast accuracy and valuable insights for informing monetary policy decisions.

Machine learning methods have also been applied. The Bank of England published a working paper testing penalised regressions and tree-based methods for inflation forecasting (Joseph et al., 2022), while Nason and Palasciano (2025) used recurrent neural networks alongside AR baselines. These approaches show promise, but gains are incremental and models are more complex to maintain and communicate.

The broad lesson, echoed in the Bernanke review (2024), is that univariate benchmarks remain difficult to beat. Complex models can add value in specific cases, but simpler approaches that are transparent, easy to update, and maintainable often perform just as well in real time. This underpins the focus here on autoregressive and UCSV models.

3 Data

This study uses monthly, non–seasonally-adjusted Consumer Prices Index (CPI) series published by the UK Office for National Statistics (ONS). The sample runs from January 1996 to June 2025 and includes the headline "All Items" CPI index together with the twelve COICOP 2-digit divisions. Examples of the divisions include: "Food and non-alcoholic beverages", "Clothing and footwear", "Housing, water, electricity, gas and other fuels", and "Transport". The full list of divisions and their corresponding basket weights is provided in the annex table 15. Basket weights are updated annually by ONS, expressed in parts per thousand and rescaled to sum to unity for aggregation.

Let $P_{i,t}$ denote the level of index i=0,…,12 in month t. Monthly inflation is defined as month-on-month percentage change:

$$\pi_{i,t} = 100\left(\frac{P_{i,t}}{P_{i,t-1}} - 1\right), t = 1 \dots, T_i$$

No seasonal adjustment or annualisation is applied prior to estimation. Instead, seasonality is modelled explicitly within the state-space framework described in Section 4. A distinction is made between the transformations used for estimation: the autoregressive benchmark is estimated on log differences of $P_{i,t}$, while the UCSV model is estimated directly on monthly inflation rates $\pi_{i,t}$. For comparability and policy relevance, however, all forecasts are converted to year-on-year (YoY) CPI inflation before evaluation. Reporting in YoY terms is consistent with the Bank of England's inflation target and facilitates comparison across divisions and with external forecasts.

Descriptive statistics

Table 1 reports the mean and standard deviation of monthly inflation rates by division across three subsamples: pre-COVID (2016–2019), COVID and energy-shock (2020–2023), and post-COVID (2024 onwards). These statistics illustrate clear differences in both the level and volatility of inflation across regimes. For example:

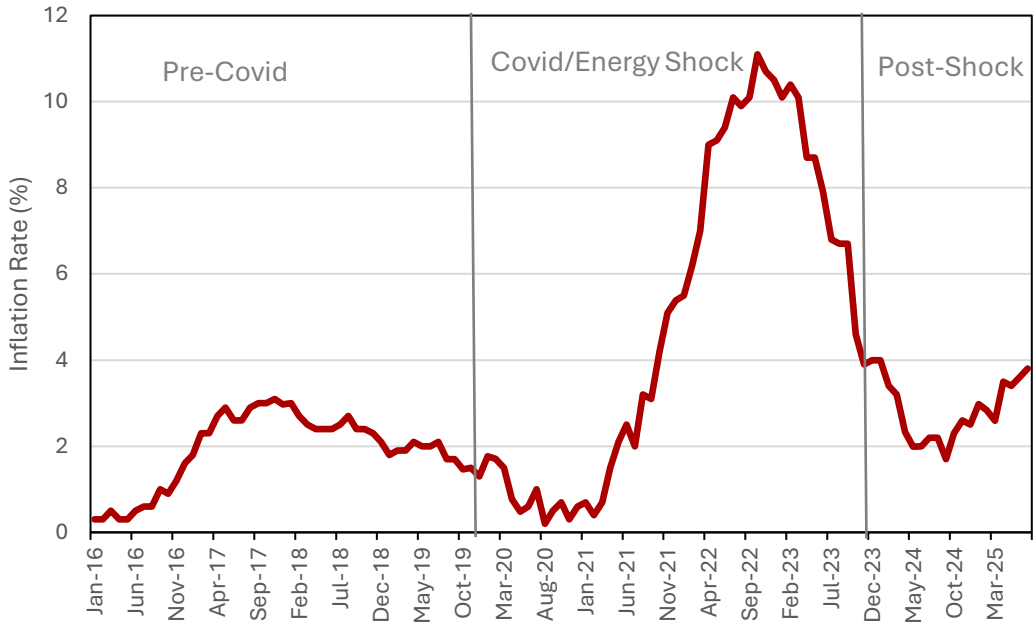
- **Pre-COVID (2016–2019):** Headline inflation was relatively stable, averaging around [X%] with low dispersion. Most divisions showed modest monthly variability.
- **COVID and energy-shock (2020–2023):** Volatility increased sharply, especially in "Food and non-alcoholic beverages" and "Housing and utilities", reflecting supply disruptions and energy price shocks. Standard deviations of monthly inflation more than doubled in these categories.
- **Post-COVID (2024–):** Volatility has moderated compared with the COVID period but remains above pre-COVID levels, particularly for services, where persistence is stronger.

Table 1 and Figure 1 report year-on-year (YoY) UK CPI inflation, while Table 2 and Figure 2 show the corresponding month-on-month (MoM) rates. These simple statistics highlight both regime shifts in the level of inflation and differences in volatility across pre-COVID (2016–2018), the COVID and energy-shock period (2020–2023), and the post-COVID phase (2024 onwards).

Table 1: Year-on-Year Headline CPI Inflation: Mean and Standard Deviation

Period	Mean (%)	Standard Deviation (pp)
Pre-COVID (2016–2019)	1.91	0.85
COVID & Energy Shock (2020–2023)	4.96	3.75
Post-COVID (2024–2025)	2.80	0.66

Figure 1: UK Headline Inflation (Year-on-Year)

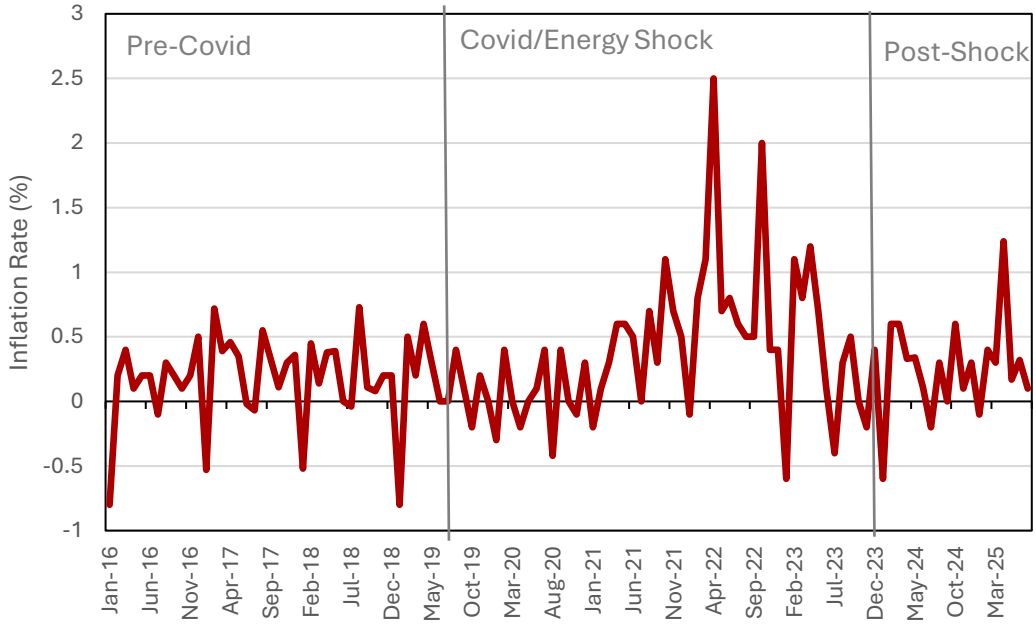


At the YoY frequency, inflation was relatively stable before COVID, averaging 1.9% with a standard deviation of 0.8 percentage points. The pandemic initially pushed inflation close to zero in 2020, but this was followed by a sharp rise to a peak of 11.1% in October 2022 – the highest in over four decades. Average inflation during the COVID and energy-shock period was almost 5.0%, with volatility more than four times its pre-COVID level. In the post-COVID period, headline inflation has moderated to an average of 2.8%, but volatility remains above pre-COVID norms, driven mainly by persistent services inflation.

Table 2. Month-on-Month Headline CPI Inflation: Mean and Standard Deviation

Period	Mean (%)	Standard Deviation (pp)
Pre-COVID (2016–2019)	0.16	0.32
COVID & Energy Shock (2020–2023)	0.41	0.57
Post-COVID (2024–2025)	0.26	0.37

Figure 2: UK Headline Inflation (Month-on-Month)

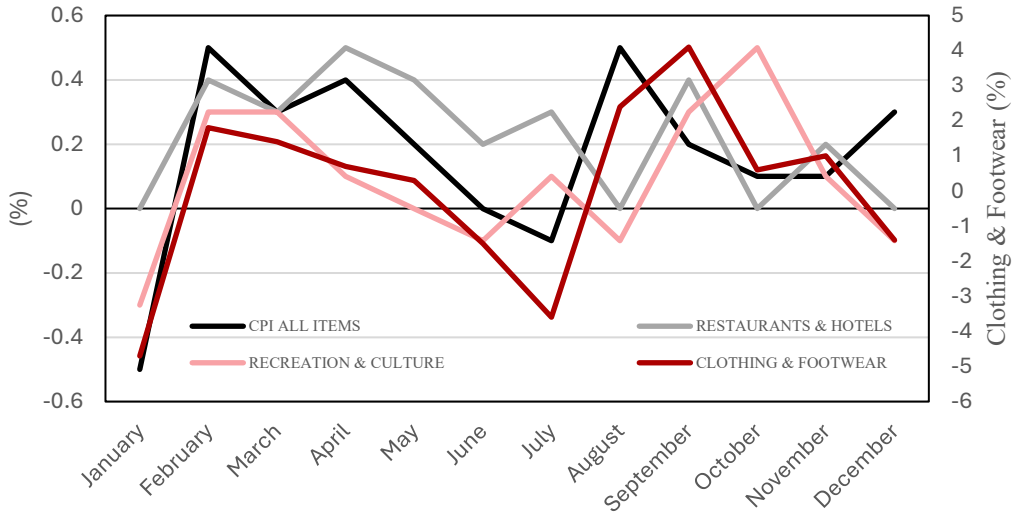


MoM statistics reveal additional features of the inflation process. Pre-COVID, monthly changes were modest (0.16% on average) with limited dispersion (0.32pp). During the COVID and energy-shock period, both the mean (0.41%) and standard deviation rose sharply. Since 2024, monthly inflation has moderated but remains more volatile than in the pre-COVID regime.

The MoM chart also makes clear the seasonal patterns in UK inflation. A sharp fall is observed each January, reflecting post-Christmas discounting, alongside a shallower but visible dip in July during summer sales. These seasonal cycles persisted through COVID, though their amplitude varied, and appear somewhat dampened in the post-COVID period. Importantly, seasonal dynamics are not uniform across CPI divisions: food and energy are dominated by supply-driven volatility, while clothing and transport display strong calendar-linked cycles. This heterogeneity underlines the importance of modelling seasonality explicitly rather than assuming uniform behaviour across components.

Figure 3 reports average month-on-month inflation over 2010–2019 for headline CPI and selected divisions. Seasonality is pronounced but heterogeneous. Clothing & footwear shows by far the strongest pattern, with large markdowns in January (≈ -5% on the right axis) and a second, smaller clearance in July (≈–3.5%), followed by partial rebounds in February–March and August–October. Recreation & culture exhibits milder seasonality: values hover near zero through most of the year, with a small dip in September and a clearer October uptick, consistent with autumn pricing of cultural goods and packaged travel. Restaurants & hotels follow a comparatively smooth profile, with modest positives in spring (March–May) and summer/early autumn and shallow dips in winter; the amplitudes are much smaller than in clothing. These contrasts highlight the heterogeneous nature of CPI seasonality across divisions, underlining the importance of modelling seasonal effects explicitly rather than assuming uniform behaviour across the basket.

Figure 3: Monthly-on-month average rate 2010-2019 (%)



4 Methodology

Unobserved components (UC) models are a class of state-space models that decompose an observed time series into latent states such as trend, cycle, seasonality, and irregular disturbances (Harvey, 1989; Blake & Mumtaz, 2017). They are widely used in macroeconomics and central banking because the state-space representation allows the latent components, such as trend and seasonality, to evolve flexibly over time and to be updated recursively as new data arrive. This framework provides a flexible and transparent way of distinguishing between persistent and transitory movements in economic data, and of modelling uncertainty in a probabilistic setting.

While UC models can be estimated using maximum likelihood and the Kalman filter (Harvey, 1989; Durbin & Koopman, 2012), this study adopts a Bayesian approach. Bayesian estimation is particularly well-suited to models with stochastic volatility (Kim, Shephard & Chib, 1998) and directly delivers full predictive distributions, which are required for density forecast evaluation

A univariate unobserved components model with stochastic volatility (UCSV) is estimated for each CPI series. The specification follows the structure proposed by Stock and Watson (2016, 2019) for inflation, extended here to monthly frequency with trigonometric seasonal components rather than fixed dummies.

4.1 Observation equation:

Let $\pi_{i,t}$ denote monthly inflation for a given CPI component i. The observation equation is:

$$\pi_{i,t} = \tau_{i,t} + \sum_{k=1}^{K} \gamma_{i,t}^{(k)} + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim N(0, \sigma_{\varepsilon,i,t}^2)$$

Where $\tau_{i,t}$ is the non-seasonal trend (local level), $\gamma_{i,t}^{(k)}$ is the seasonal component and $\epsilon_{i,t}$ is the irregular disturbance with time-varying variance.

4.2 State Equations:

The **trend** is modelled as a local level that follows a random walk:

$$\tau_{i,t} = \tau_{i,t-1} + \eta_{i,t}, \quad \eta_{i,t} \sim N(0, \sigma^2_{\eta,i,t})$$

This specification treats the underlying inflation trend as persistent but time-varying. Shocks to the trend are permanent, allowing the model to capture shifts in long-run inflation dynamics without requiring arbitrary structural breaks. In practice, this means the model can accommodate gradual disinflation episodes (e.g. 2012-2014) as well as persistent increases in trend inflation during shocks such as the post-pandemic energy crisis.

Seasonality refers to systematic within-year patterns that repeat regularly across months. In CPI data, such effects are economically intuitive: for instance, clothing and footwear prices typically fall in January and July during winter and summer sales, while airfares and package holidays peak in school holiday months. Capturing these predictable intra-year cycles is crucial for distinguishing temporary fluctuations from underlying inflation trends.

Here, seasonality is represented by a trigonometric formulation with *K* harmonics (Harvey, 1989). For each harmonic k = 1,…,K with frequency $\lambda_k = \frac{2\pi k}{12}$. Each harmonic introduces a pair of sine and cosine states with frequency.

$$\gamma_{i,t}^{(k)} = \gamma_{i,t-1}^{(k)} \cos(\lambda_k) + \gamma_{i,t-1}^{*(k)} \sin(\lambda_k) + \xi_{i,t}^{(k)}$$

$$\gamma_{i,t}^{*(k)} = -\gamma_{i,t-1}^{(k)} \sin(\lambda_k) + \gamma_{i,t-1}^{*(k)} \cos(\lambda_k) + \xi_{i,t}^{*(k)}$$

The innovations of the seasonal components ($\xi_{i,t}^{(k)}$ and $\xi_{i,t}^{*(k)}$)` are jointly Gaussian,

$$\begin{bmatrix} \xi_{i,t}^{(k)} \\ \xi_{i,t}^{*(k)} \end{bmatrix} \sim N(0, \sigma^2_{s,i,t} I_2),$$

where both the sine and cosine components share a common time-varying variance $\sigma^2_{s,i,t}$. The overall seasonal component is the sum across the harmonics:

$$\gamma_{i,t} = \sum_{k=1}^{K} \gamma_{i,t}^{k}$$

This trigonometric structure has two advantages: (i) it captures smooth and cyclical seasonal effects without imposing fixed monthly dummies, and (ii) the stochastic specification allows the amplitude of seasonal fluctuations to evolve over time, which is important when sale patterns or travel demand shift in response to structural changes.

4.3 Stochastic volatility specification

The variances of the innovations to the trend, irregular, and seasonal components are themselves time-varying, evolving according to log-AR(1) processes:

$$log\ \sigma^2_{m,i,t} = \mu_m + \phi_m\left(log\ \sigma^2_{m,i,t-1} - \mu_m\right) + \upsilon_{m,i,t}, \qquad \upsilon_{m,i,t} \sim N(0, \sigma^2_{\upsilon,m})$$

For $m \in \{\eta, \varepsilon, s\}$, denoting the trend, irregular, and seasonal components respectively.

Allowing for stochastic volatility (SV) is motivated by the well-documented heteroskedasticity of inflation and other macroeconomic time series. From a methodological perspective, Bayesian SV models are now standard for handling time-varying uncertainty, following Kim, Shephard and Chib (1998) and Omori et al. (2007). These frameworks provide efficient likelihood-based inference, overcoming the limitations of constant-variance or ARCH-type specifications.

In the context of inflation, SV plays a particularly important role. Stock and Watson (2016, 2019) highlight that the variance of both permanent and transitory components shifts considerably across regimes: for example, volatility was muted in the "Great Moderation" but surged during the global financial crisis and pandemic. Incorporating SV allows the model to:

1. Distinguish more clearly between persistent and transitory shocks when volatility rises.
2. Adjust the width of predictive densities in line with prevailing uncertainty, yielding forecast "cones" that naturally widen in turbulent periods and narrow in stable regimes.
3. Capture changes in the amplitude of seasonal effects, such as more pronounced January and July clothing sales during some years than others.

Together, these features ensure that the model remains flexible across macroeconomic environments, producing more realistic and better-calibrated forecast densities than constant-variance alternatives.

4.4 Estimation

Each series is estimated separately using Bayesian Markov chain Monte Carlo (MCMC) methods. The sampler alternates between:

1. Sampling SV parameters and latent log-variances for trend, irregular, and seasonal components.
2. Drawing the full state vector ($\tau_{i,t}, \gamma_{i,t}$) via the Carter–Kohn simulation smoother.
3. Iterating until convergence, with an initial burn-in followed by retained posterior draws.

Forecasts for horizons h= 1,…,6 months are generated by simulating the state-space model forward for each posterior draw, yielding a full predictive distribution.

The estimation was implemented in *R*, using the *stochvol* package (Kastner, 2016) for stochastic volatility sampling. A custom Carter–Kohn simulation smoother was utilised to draw the state vector, as no suitable *R* package was available for Bayesian estimation of unobserved components models.

4.5 Bottom-up aggregation

In the bottom-up approach, division-level forecasts are re-aggregated to form headline CPI. For evaluation date t, contemporaneous CPI weights $w_{i,t}$ are obtained from the ONS basket corresponding to the appropriate year, rescaled to sum to unity across divisions $\mathcal{D}$:

$$\sum_{i \in \mathcal{D}} w_{i,t} = 1$$

Since the ONS updates CPI weights annually to reflect changing consumer spending patterns, using year-specific weights ensures that the aggregation accurately captures the relative importance of each division at the time of the forecast

Let $f_{i,t+h}^{(j)}$ be the forecast draw for component i, horizon h, and posterior draw index j. For each draw j, the aggregated forecast is computed as:

$$f_{t+h}^{(j)} = \sum_{i \in \mathcal{D}} w_{i,t} \, f_{i,t+h}^{(j)}$$

This "draw-then-aggregate" procedure preserves the coherence of simulated forecast paths. Headline point forecasts are taken as the posterior median across draws; densities are given by the full empirical distribution.

4.6 Top-down comparison

For comparison, the UCSV model is also estimated directly on the headline CPI series. Forecast evaluation compares top-down and bottom-up forecasts against benchmarks.

4.7 Forecast evaluation and benchmarks

Forecast evaluation is conducted in a recursive pseudo-real-time design with an expanding window. Both point and density accuracy are assessed. A density forecast is an estimate of the full predictive distribution of a variable, not just its expected value, and thus captures uncertainty as well as central tendency (Diebold, Gunther & Tay, 1998; Amisano & Giacomini, 2007). This distinction is crucial for policymakers, since it allows forecasts to convey both central projections and the likelihood of tail events. In line with recent recommendations (Bernanke, 2024), this study therefore evaluates models not only on point accuracy but also on their ability to generate well-calibrated and sharp predictive densities.

- **Point forecasts.** Root mean squared error (RMSE) is reported at horizons of one to six months. Statistical significance of forecast differences is assessed using the Diebold–Mariano (DM) test (Diebold & Mariano, 1995), which compares average squared error losses between models. A full technical formulation of the DM test is provided in the annex.
- **Density forecasts**: Beyond point accuracy, it is important to evaluate how well models capture the full predictive distribution of inflation. Three complementary metrics are used:

    1. **Log predictive score (LS).** A proper scoring rule that evaluates the probability density assigned to the realised outcome. Higher scores indicate models that place greater mass on the observed value, rewarding sharp and well-centred forecasts. LS has become a standard in density forecast evaluation (Amisano & Giacomini, 2007).

    2. **Continuous ranked probability score (CRPS).** A measure of the distance between the forecast distribution and the realised outcome across the full support. Unlike LS, which focuses on the realised point, CRPS evaluates both sharpness (narrowness of the distribution) and calibration (alignment with realised outcomes). Lower CRPS values indicate better density forecasts. Alongside raw values, a **CRPS skill score** is reported to facilitate interpretation across horizons. This is defined as one minus the ratio of the model's CRPS to that of the benchmark AR, so that positive values indicate improvement over AR and negative values indicate deterioration. See Gneiting & Raftery (2007) for background on CRPS and skill scores.

    3. **Coverage and calibration tests.** These assess whether empirical outcomes fall within nominal prediction intervals (e.g. 50% or 90%) at the expected frequencies. In addition, probability integral transform (PIT) histograms are used as a diagnostic of calibration, with uniformity indicating well-calibrated densities (Diebold, Gunther & Tay, 1998; Gneiting & Katzfuss, 2014).

Together these measures provide a multi-dimensional evaluation: LS rewards models that assign higher likelihood to realised values, CRPS balances sharpness and calibration while allowing for relative skill comparisons, and coverage tests directly assess interval reliability. This ensures that both the centre and the tails of the predictive distribution are evaluated, addressing recent calls by The Bernanke Review (2024) to place greater emphasis on uncertainty in inflation forecasting. Technical definitions and formulae for LS, CRPS, CRPS skill, coverage statistics, and PIT diagnostics are set out in the annex.

In line with the inflation forecasting literature, this paper uses a simple autoregressive (AR) model as the benchmark model. AR models assume that the current value of a series can be explained by a linear combination of its past values and an error term. They are widely used because they are

transparent, parsimonious, and often difficult to outperform in real-time forecasting (Stock & Watson, 2007). Formally, an AR(p) process is written as:

$$y_t = c + \Sigma_{i=1}^{p} \Phi_i y_{t-i} + \varepsilon_t$$

The optimal lag length is chosen using the Bayesian Information Criterion (BIC; Schwarz, 1978), which favours more parsimonious specifications and guards against overfitting in recursive samples.

5. Results

This section evaluates the forecasting performance of the UCSV model relative to a benchmark AR. Results are presented separately for point forecasts and density forecasts. Within UCSV, we distinguish between two approaches: top-down (UCSV-TD), where the model is estimated directly on headline CPI, and bottom-up (UCSV-BU), where forecasts are generated for disaggregated CPI components and then aggregated to headline. Unless otherwise stated, all UCSV results reported in this section are based on the trigonometric seasonal specification with two harmonics. This choice follows robustness testing (see Section X), where alternative specifications with three harmonics were considered.

5.1 Point Forecast Accuracy

Pre-Covid (Jan 2016 -Dec 2019)

At the one-month horizon, the AR benchmark was the most accurate, with both UCSV specifications producing larger errors. From horizon 2 onwards, however, UCSV consistently outperformed AR, especially in the top-down specification (Table 3). Relative to AR, top-down UCSV reduced RMSE by 40–60% at horizons 3–6, with significance confirmed by Diebold–Mariano tests (Table 4). Bottom-up forecasts also improved on AR from horizon two onwards, though gains were smaller and less robust.

Table 3. RMSE of Point Forecasts (Pre-COVID)

Horizon	AR (Benchmark)	UCSV (Top-down)	UCSV (Bottom-up)
1	0.18	0.36	0.33
2	0.51	0.42	0.43
3	0.77	0.43	0.50
4	1.03	0.53	0.63
5	1.32	0.64	0.78
6	1.61	0.67	0.88

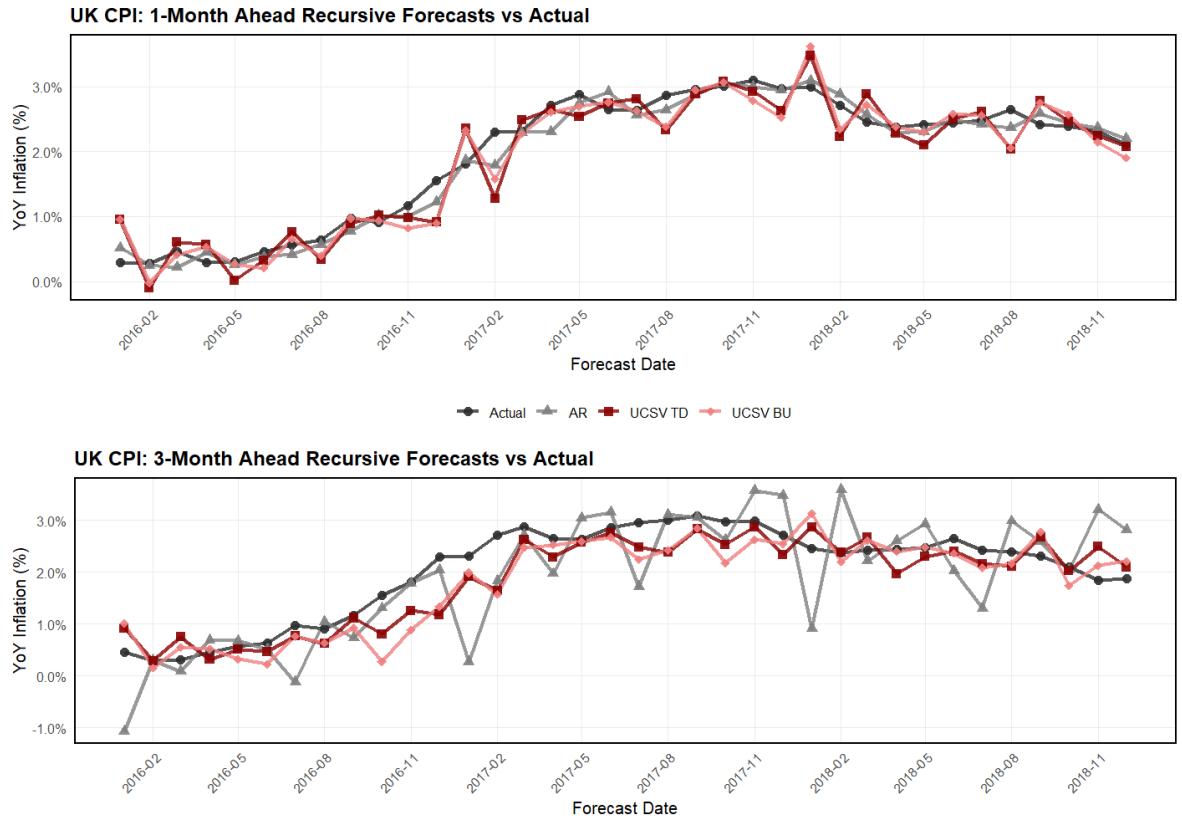
Table 4. Relative RMSE vs AR (Pre-COVID)

Horizon	UCSV (Top-down)	UCSV (Bottom-up)
1	2.00	1.82
2	0.81	0.83
3	0.55**	0.64*
4	0.51**	0.62*
5	0.48**	0.60*
6	0.41***	0.54**

Notes: Values < 1 indicate UCSV outperforms AR. Stars indicate Diebold–Mariano significance vs AR († p<0.10, * p<0.05, ** p<0.01, *** p<0.001).

Figure 4 illustrates these patterns. At the one-month horizon, all models tracked headline inflation closely, which remained in a narrow 0–3% range. However, UCSV forecasts display some sharp spikes, particularly in 2017, which are examined further in the Discussion section when considering seasonal specification. At the three-month horizon, AR forecasts became more volatile, especially during 2018, while UCSV-TD maintained steadier tracking. These visual patterns reinforce the RMSE results reported in Tables 3 and 4.

Figure 4:



COVID / Energy Spike Period (Jan 2020 – Dec 2023)

During the COVID and energy-price shock, forecasting errors rose sharply across all models (Table 5), reflecting the unprecedented volatility in inflation dynamics relative to the pre-COVID period. The sharp swings in energy prices, supply-chain disruptions, and policy interventions created conditions where simple autoregressive benchmarks struggled to capture turning points. At the one-month horizon, all three models performed similarly, with no significant differences, highlighting the difficulty of short-term prediction in such a turbulent environment. From h=2 onwards, however, both UCSV variants consistently outperformed AR, with the top-down specification delivering the clearest gains. Relative RMSEs (Table 4) show that top-down UCSV reduced errors by around 30% at horizons 4–6, with statistical significance confirmed by DM tests. The bottom-up approach also improved on AR at shorter horizons, but its advantage diminished beyond h=4, where forecast errors widened again, likely reflecting the amplification of volatility when aggregating disaggregated shocks during this period.

Table 5. RMSE of Point Forecasts (COVID / Energy Spike)

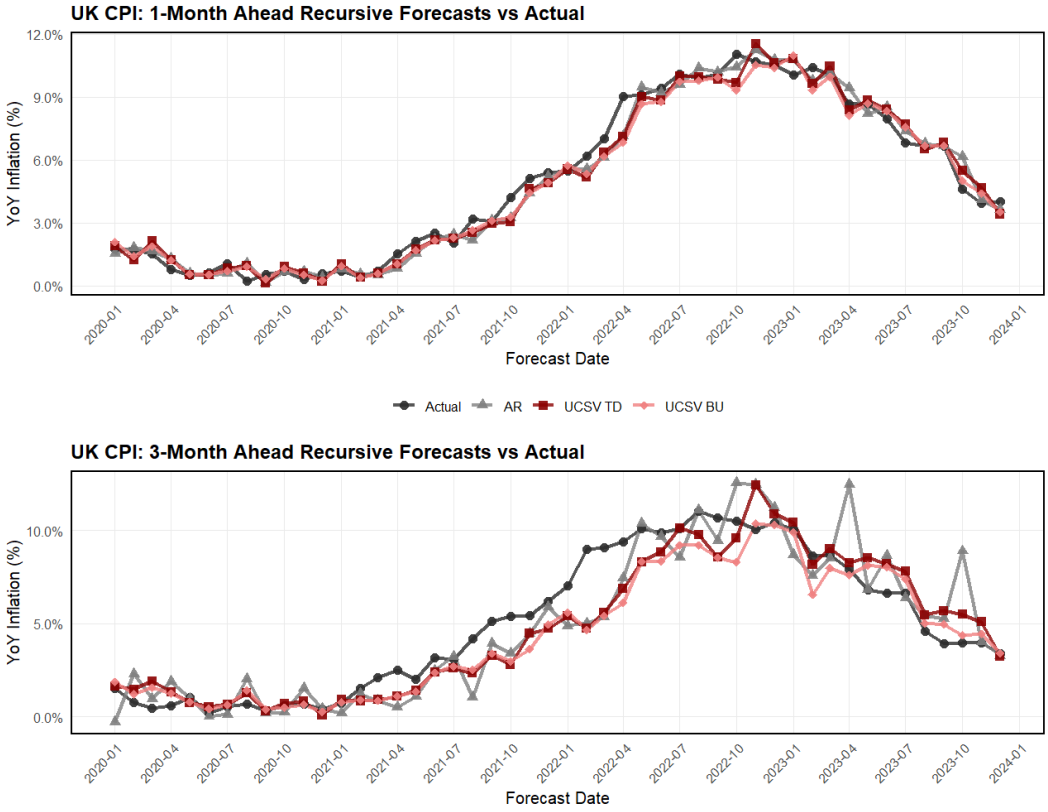
Horizon	AR (Benchmark)	UCSV (Top-down)	UCSV (Bottom-up)
1	0.59	0.60	0.61
2	1.15	1.02	1.02
3	1.71	1.42	1.44
4	2.50	1.74	1.82
5	3.10	2.14	2.32
6	3.72	2.58	3.16

Table 6. Relative RMSE vs AR (COVID / Energy Spike)

Horizon	UCSV (Top-down)	UCSV (Bottom-up)
1	1.02	1.04
2	0.89	0.89
3	0.83	0.84
4	0.70*	0.73†
5	0.69*	0.75
6	0.69*	0.85

Figure 5 captures the turbulent COVID and energy shock episode, with inflation swinging from near-zero to over 10% before returning towards the target. At the one-month horizon, all models keep pace with the initial surge, though differences become clearer during the disinflation phase. At three months ahead, AR shows large swings and occasional missed turning points, while UCSV top-down provides more stable directional guidance. This aligns with Table 6, where UCSV achieved sizeable RMSE gains relative to AR, despite elevated errors across all models in this volatile environment.

Figure 5



Post-Covid (Jan 2024 - June 2025)

Forecast performance shifted again in the post-COVID/Energy shock period (Table 7). At h=1, AR retained a slight advantage, though differences were not statistically significant. From h=2 onwards, both UCSV variants decisively outperformed AR, with the strongest gains delivered by the top-down specification. Relative RMSEs (Table 8) show that UCSV-TD reduced errors to less than one-third of AR by h=6. UCSV-BU was competitive at horizons 2–3, even marginally outperforming UCSV-TD, but weakened again at longer horizons.

Table 7. RMSE of Point Forecasts (Post-COVID/Energy shock)

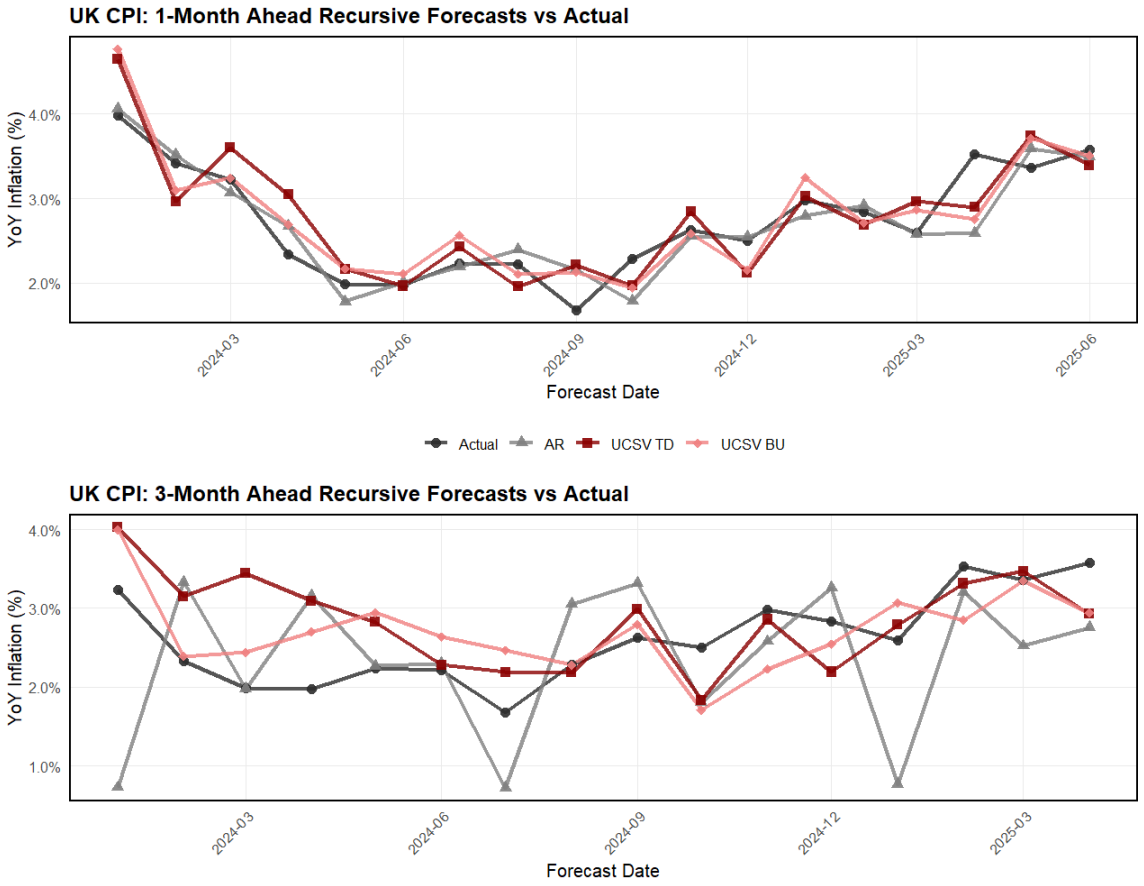
Horizon	AR (Benchmark)	UCSV (Top-down)	UCSV (Bottom-up)
1	0.32	0.39	0.36
2	0.64	0.44	0.39
3	1.01	0.65	0.56
4	1.67	0.77	0.74
5	2.22	0.93	1.05
6	2.97	1.00	2.13

Table 8. Relative RMSE vs AR (Post-COVID//Energy shock)

Horizon	UCSV (Top-down)	UCSV (Bottom-up)

1	1.27	1.18
2	0.70	0.61*
3	0.65	0.55†
4	0.46*	0.44*
5	0.42**	0.47*
6	0.34**	0.72

Figure 6



5.2 Density Forecast Accuracy

Evaluation of density forecasts provides a more comprehensive assessment of predictive performance than point forecasts alone, as it considers both the calibration and sharpness of the entire predictive distribution. Two complementary metrics are reported: the mean log score (LS), where higher values indicate greater likelihood assigned to realised outcomes, and the continuous ranked probability score (CRPS), where lower values denote superior calibration and sharpness.

Pre-Covid (Jan 2016 -Dec 2019)

In the relatively stable pre-pandemic period, the autoregressive benchmark performed slightly better at the one-month horizon, but from the two-month horizon onwards both UCSV specifications improved materially on AR. The top-down variant delivered the largest gains, with LS differences exceeding +1.7 by the six-month horizon and improvements statistically significant from horizon

three onwards. The bottom-up model also outperformed AR at intermediate horizons, though its advantage diminished at longer horizons. CRPS results corroborate this pattern, showing steadily rising skill scores for UCSV-TD and weaker, less consistent gains for UCSV-BU.

Table 9. Mean Log Scores (Pre-COVID)

Horizon	AR (Benchmark)	UCSV (Top-down)	UCSV (Bottom-up)	ΔLS TD–AR	ΔLS BU–AR
1	0.24	–0.41	–0.36	–0.65	–0.60
2	–0.86	–0.72	–0.64	+0.14	+0.22
3	–1.55	–0.90†	–0.85†	+0.65	+0.71
4	–2.06	–1.05*	–1.19*	+1.01	+0.97
5	–2.39	–1.18**	–1.31*	+1.21	+1.07
6	–3.00	–1.26***	–1.55***	+1.74	+1.46

Notes: Higher = better. Stars indicate Amisano–Giacomini test significance vs AR.

Table 10. CRPS and CRPS Skill (Pre-COVID)

Horizon	CRPS (AR)	CRPS (UCSV TD)	CRPS (UCSV BU)	CRPS Skill TD	CRPS Skill BU
1	0.10	0.20	0.19	–0.97	–0.89
2	0.29	0.25	0.27	+0.13	+0.07
3	0.44	0.28	0.41	+0.36	+0.07
4	0.59	0.33	0.70	+0.43	–0.19
5	0.74	0.39	1.10	+0.47	–0.49
6	0.92	0.41	1.53	+0.55	–0.66

Notes: Lower CRPS = better. Skill = 1 – CRPS(UCSV)/CRPS(AR); positive = UCSV improvement.

COVID / Energy Spike Period (Jan 2020 – Dec 2023)

The pandemic and subsequent energy-price shock created exceptional volatility, raising the difficulty of producing well-calibrated densities. At the one-month horizon, the AR benchmark retained an advantage, reflecting sharper but ultimately mis-calibrated distributions. From the two-month horizon onwards, however, UCSV models, particularly UCSV-TD, assigned markedly higher likelihood to realised outcomes. Log score gains exceeded +2.9 by the six-month horizon, with improvements statistically significant at all horizons beyond one month. CRPS results reinforce this conclusion: UCSV-TD reduced forecast distance by around 30–35% at longer horizons, while UCSV-BU delivered short-horizon gains but deteriorated when volatility from disaggregated shocks propagated through aggregation.

Table 11. Mean Log Scores (COVID / Energy Spike)

Horizon	AR (Benchmark)	UCSV (Top-down)	UCSV (Bottom-up)	ΔLS TD–AR	ΔLS BU–AR
1	–1.53	–0.97*	–1.14	+0.55	+0.39
2	–2.96	–1.48***	–1.70**	+1.49	+1.26
3	–3.64	–1.89***	–2.17**	+1.74	+1.46

4	–4.26	–2.35***	–2.38***	+1.91	+1.88
5	–5.13	–2.68***	–2.69***	+2.45	+2.44
6	–5.90	–2.96***	–2.87***	+2.94	+3.03

Table 12. CRPS and CRPS Skill (COVID / Energy Spike)

Horizon	CRPS (AR)	CRPS (UCSV TD)	CRPS (UCSV BU)	CRPS Skill TD	CRPS Skill BU
1	0.35	0.33	0.36	+0.05	–0.02
2	0.70	0.56	0.66	+0.19	+0.05
3	1.04	0.80	1.14	+0.23	–0.09
4	1.52	1.02	1.84	+0.33	–0.21
5	1.97	1.28	2.66	+0.35	–0.35
6	2.41	1.57	3.59	+0.35	–0.49

Post-Covid / Energy Shock (Jan 2024-June 2025)

In the most recent regime, the same broad ranking emerged. At the one-month horizon, AR densities were sharper and slightly better aligned with outcomes. From horizon two onwards, however, UCSV-TD delivered consistent and substantial gains, with LS improvements exceeding +3.0 and CRPS skill reaching +0.69 by the six-month horizon. UCSV-BU was competitive at horizons two and three, occasionally matching UCSV-TD, but weakened thereafter as densities became increasingly dispersed.

Table 13. Mean Log Scores (Post-COVID / Energy Shock)

Horizon	AR (Benchmark)	UCSV (Top-down)	UCSV (Bottom-up)	ΔLS TD–AR	ΔLS BU–AR
1	–0.22	–0.56	–0.59	–0.34	–0.37
2	–1.15	–0.93	–0.82	+0.22	+0.34
3	–1.63	–1.23	–1.16	+0.40	+0.47
4	–2.79	–1.40*	–1.60†	+1.39	+1.20
5	–3.63	–1.57*	–1.93†	+2.06	+1.70
6	–4.77	–1.68**	–2.00*	+3.09	+2.77

Table 14. CRPS and CRPS Skill (Post-COVID / Energy Shock)

Horizon	CRPS (AR)	CRPS (UCSV TD)	CRPS (UCSV BU)	CRPS Skill TD	CRPS Skill BU
1	0.16	0.23	0.24	–0.45	–0.50
2	0.39	0.29	0.39	+0.25	+0.01

3	0.60	0.41	0.86	+0.32	–0.44
4	1.00	0.48	1.62	+0.52	–0.61
5	1.42	0.58	2.53	+0.59	–0.78
6	2.04	0.64	3.01	+0.69	–0.47

5.3 Results Summary

Across all three periods, the AR benchmark remained broadly competitive at the one-month horizon, but differences with UCSV were small and rarely statistically significant. From horizon two onwards, however, the UCSV framework consistently outperformed the benchmark, with the top-down specification providing the most reliable gains. Before the pandemic, UCSV-TD cut RMSE by 40–60% at horizons three to six, while density metrics showed significant improvements in both sharpness and calibration. The COVID and energy-price spike saw a general rise in forecast errors, yet the relative ranking between models held: UCSV-TD continued to deliver sizeable and statistically significant gains, whereas UCSV-BU was only effective at short horizons and deteriorated at longer ones, producing over-dispersed densities. In the post-COVID sample, both UCSV variants decisively outperformed AR from horizon two onwards, with UCSV-TD reducing RMSE by two-thirds and achieving the largest density gains, while BU was competitive only at shorter horizons.

Taken together, these results show that UCSV-TD is a robustly superior forecasting method, delivering consistent improvements in both point and density accuracy across different macroeconomic regimes. The bottom-up variant benefits from disaggregated information and can improve short-term accuracy, but its long-horizon densities are less reliable. The autoregressive benchmark remains difficult to beat at h=1, but quickly loses ground, underscoring the value of richer state-space methods for medium-term inflation forecasting. Robustness checks with alternative seasonal specifications do not overturn these conclusions.

6. Robustness Checks

6.1 UCSV Seasonality Specification:

A key robustness check concerns the number of harmonics in the trigonometric seasonal component. Harmonics represent cycles of different frequencies: the first captures the broad annual cycle, while higher harmonics allow for semi-annual or finer seasonal variation (Harvey, 1989).

The baseline specification employs two harmonics. To test sensitivity, the model was re-estimated with three harmonics. Tables 18 and 19 (in the annex) report pre-COVID results. The choice of seasonal specification makes very little difference. For UCSV-TD, RMSE and relative RMSE are essentially unchanged across horizons. For UCSV-BU, the three-harmonic variant yields slightly smaller errors at horizons 2–3, but the improvements are marginal and not consistent across the forecast horizon. Importantly, the relative ranking of models is unaffected: UCSV-TD remains the strongest performer, UCSV-BU provides moderate gains at short horizons, and AR lags behind from horizon 2 onwards.

Although additional harmonics could, in principle, allow for richer seasonal dynamics, exploratory tests with more than three harmonics (not reported) confirmed no further benefits and in some cases slightly degraded performance. This likely reflects over-parameterisation relative to the short evaluation samples. Overall, two harmonics provide a parsimonious and effective seasonal specification, and this choice is retained for the main results.

6.2 Prior Specification Sensitivity Analysis

Prior specification represents a potential source of sensitivity in Bayesian estimation, particularly when sample sizes are modest or structural breaks limit the effective sample available for parameter learning (Koop and Potter, 2007). In the UCSV framework, priors govern initial beliefs about the volatility of trend, seasonal, and irregular innovations, as well as the dynamics of the stochastic volatility processes. While weakly informative priors are generally preferred to allow data to dominate posterior inference, hyperparameter choices may still influence results when the likelihood provides limited information.

To assess robustness, alternative prior configurations were tested representing tighter and looser beliefs about volatility parameters relative to a moderate baseline specification. Post-COVID recursive estimation was conducted under each configuration to quantify sensitivity of forecast performance. RMSE differences between forecasts generated under alternative and baseline priors range from 0.06 to 0.12 percentage points across forecast horizons (Annex Table 20). These differences represent modest variation. While the largest differences occur at medium horizons (2-4 months), no systematic bias emerges across configurations.

These results indicate that posterior inference is robust to reasonable variations in prior specification. The relatively small forecast differences suggest the data samples provides sufficient information for the likelihood to dominate prior beliefs, consistent with the theoretical expectation that informative data should overwhelm weak priors in large samples (Bernardo and Smith, 2000). The main results therefore appear robust to alternative prior parameterisations.

6.3 HP Filter Smoothness Parameter Sensitivity Test

The initial trend states were set using a Hodrick–Prescott (HP) filter. Following Ravn and Uhlig (2002), the theoretically appropriate smoothing parameter (λ) for monthly data is $\lambda = 129{,}600$, scaled from the quarterly benchmark of $\lambda = 1{,}600$. In practice, however, alternative choices are often used, such as $\lambda = 14{,}400$ or even re-using $\lambda = 1{,}600$ for monthly series. To ensure robustness, forecasts were re-estimated with all three values. While $\lambda = 129{,}600$ provides the most consistent theoretical foundation, the main results reported in results section are based on $\lambda = 14{,}400$, which lies between the quarterly and monthly settings and is frequently adopted in applied work. Importantly, the choice of λ made little material difference: RMSEs, log scores and relative model rankings were virtually unchanged across the three calibrations. This reflects the fact that the HP filter is only used to initialise the trend component, with subsequent state updates driven by the data and stochastic volatility dynamics.

6.4 Density Calibration and Coverage Tests

Coverage tests and PIT histograms (see annex table 16 and 17) provide an additional diagnostic of forecast calibration, complementing log score and CRPS measures reported in the Results section. Across the three subsamples, UCSV-TD densities were generally well-calibrated: PIT means clustered close to 0.5, interval coverage rates aligned with nominal levels, and PIT distributions approximated uniformity. These findings indicate that UCSV-TD successfully allocated probability mass around realised outcomes, consistent with its superior log score performance.

By contrast, UCSV-BU tended to generate over-dispersed densities at longer horizons. This is evident in coverage tests, where empirical coverage exceeded nominal rates, and in PIT histograms, which displayed a mild U-shaped pattern. These results are consistent with the CRPS evidence in Section 4, where UCSV-BU deteriorated at longer horizons.

The AR benchmark was sharper at short horizons, but PIT tests reveal occasional misallocation of probability mass in volatile periods, particularly during the COVID and energy shock episode. Taken together, these diagnostics reinforce the main conclusion that UCSV-TD provides the most reliable

density forecasts across different macroeconomic regimes, while UCSV-BU is competitive only at short horizons.

7. Discussion

The results demonstrate that the UCSV model systematically improve the forecasting of UK inflation compared with an AR benchmark, particularly at horizons beyond one month. While the AR model remains competitive at the very short horizon, consistent with the broader literature showing that near-term inflation is difficult to predict (Stock and Watson, 2007; 2010). the UCSV model's ability to adapt to drifting means and changing volatility enables it to deliver more accurate forecasts from two months onwards. The top-down specification is especially effective, producing substantial and statistically significant reductions in forecast errors and sharper, better-calibrated densities across all three regimes. These findings directly address the motivation set out in the introduction: to evaluate whether more flexible UCSV models can strengthen short-run inflation forecasting in light of recent failures, particularly during the energy-price spike highlighted in Bernanke's (2024) review.

A key contribution of this study is the systematic evaluation of density forecasts, in line with Bernanke's recommendation that policy institutions should place greater emphasis on uncertainty. The top-down UCSV consistently assigns higher probability to realised outcomes and produces narrower and more reliable predictive distributions than the autoregressive benchmark, especially during turbulent periods such as COVID-19 and the subsequent energy shocks. This is particularly valuable for policy institutions, which must not only generate central projections but also communicate the risks around them. By contrast, the AR densities tend to appear sharper at the one-month horizon but are poorly calibrated, underestimating tail risks and thereby understating the probability of extreme inflation outcomes.

The evidence on bottom-up forecasting is more mixed. Disaggregated forecasts do provide some benefits, often outperforming the autoregressive benchmark. However, these gains are small, not robust across regimes, and disappear at longer horizons, where the volatility inherent in components propagates upwards and produces over-dispersed densities. In other words, bottom-up aggregation improves modestly on the AR benchmark, but it does not outperform the top-down UCSV and cannot be considered a consistently superior approach. This finding mirrors Hubrich (2005) who show that the benefits of disaggregation are at best mixed.. For UK CPI divisions, the heterogeneity of shocks, ranging from food and energy, appears to limit the scope for bottom-up improvement. An important extension would be to test bottom-up methods at lower levels of disaggregation (e.g. COICOP 3-digit or item level), where sector-specific shocks may be more informative and aggregation could filter noise more effectively.

Despite these advantages, UCSV has clear limitations. The model does not dominate at the one-month horizon, which is often the most policy-relevant for central banks. For example, the Bank of England reports one-month-ahead RMSEs of roughly 0.17 percentage points in its Monetary Policy Report (see Nason and Palasciano, 2025), a benchmark that purely statistical models rarely surpass. This partly reflects the use of human judgement, which incorporates known policy measures such as changes to Ofgem's energy price cap, information that statistical models cannot anticipate. Moreover, UCSV is a univariate framework: it cannot generate conditional or scenario forecasts, such as the impact of an exchange-rate depreciation, which Bernanke (2024) identifies as a central requirement for modern forecasting. Multivariate approaches such as BVARs are better suited to this task, combining structural interpretability with the ability to simulate policy counterfactuals. UCSV should therefore be viewed as a complement rather than a substitute, offering a parsimonious and adaptive baseline that can be cross-checked against richer structural tools and expert judgement.

Finally, the results reinforce the case for a diversified forecasting toolkit. No single model dominates across all horizons or regimes. Combining UCSV with other approaches could reduce variance and hedge against misspecification, consistent with the forecast-combination literature (Bates and Granger, 1969). In practice, central banks already employ suites of models, ranging from UC specifications to VARs, DSGEs, and machine-learning approaches, and weigh them against structured expert judgement. This study suggests that UCSV should form part of such a toolkit, valued for its interpretability and density calibration, but not relied upon in isolation.

8. Conclusion

This paper has provided a systematic evaluation of UCSV models for UK inflation forecasting, benchmarked against AR models, with attention to both point and density performance. The results show that UCSV-TD consistently outperforms AR beyond one month, delivering lower RMSEs and superior density forecasts across pre-COVID, COVID/energy-shock, and post-COVID regimes. These findings support the case that more flexible state-space models can strengthen short-run inflation forecasting, particularly in periods of heightened volatility when standard benchmarks struggle.

The evidence on UCSV-BU is more nuanced. While it generally performs better than AR at shorter horizons, the gains are modest and not robust across regimes, and at longer horizons densities become less reliable compared with UCSV-TD. This suggests that forecasting at the division level and re-aggregating to headline does not systematically improve forecast performance for UK CPI and may even propagate volatility from heterogeneous components such as food and energy. Future work could extend the analysis to more granular data (e.g. COICOP 3-digit or item level), where disaggregation may help to extract sectoral signals while allowing aggregation to filter idiosyncratic noise more effectively.

Several directions for further research are clear.

1. First, the incremental role of stochastic volatility could be quantified by re-estimating UC models without SV, to assess how much of the improvement comes from variance adaptation versus trend flexibility.
2. Second, forecast combinations across AR, UCSV, BVARs, and machine-learning approaches could be explored as a hedge against misspecification, consistent with the forecast-combination literature.
3. Third, explicit outlier detection, as in Stock and Watson (2016), would help improve robustness to temporary price spikes. This is because large, one-off shocks in components such as airfares, energy, or package holidays can distort estimates of the underlying inflation trend if they are absorbed into the permanent component. Allowing for heavy-tailed innovations in the irregular part of the model provides a systematic way to down-weight these transitory jumps in real time, rather than relying solely on judgement. Similar modifications are already used in central banks' toolkits, where filtering out extreme but temporary shocks is seen as essential for extracting underlying inflation.
4. Finally, statistical models cannot substitute for expert judgement. Known policy changes such as Ofgem's energy price cap adjustments or fiscal measures like VAT changes on alcohol duties remain outside the reach of purely statistical approaches. The most effective approach is therefore hybrid: UCSV provides a transparent and adaptive statistical baseline, while forecasters apply structured judgement to integrate real-time policy information. Taken together, this ensures forecasts that are both analytically robust and policy-relevant, consistent with Bernanke's (2024) call for a modernised toolkit that balances technical innovation with clear communication of uncertainty.

References:

Amisano, G. and Giacomini, R. (2007) 'Comparing density forecasts via weighted likelihood ratio tests', Journal of Business & Economic Statistics, 25(2), pp. 177-190.

Atkeson, A. and Ohanian, L. E. (2001) 'Are Phillips curves useful for forecasting inflation?', Federal Reserve Bank of Minneapolis Quarterly Review, 25(1), pp. 2-11. Available at: https://www.minneapolisfed.org/research/qr/qr2511.pdf (Accessed: 7 September 2025).

Bates, J. M. and Granger, C. W. J. (1969) 'The combination of forecasts', Journal of the Operational Research Society, 20(4), pp. 451-468..

Bernanke, B. S. (2024) Review of the Bank of England's forecasting capability: Report and recommendations. London: Bank of England. Available at: https://www.bankofengland.co.uk/independent-evaluation-office/forecasting-for-monetary-policy-making-and-communication-at-the-bank-of-england-a-review

Bermingham, C. and D'Agostino, A. (2011) 'Understanding and forecasting aggregate and disaggregate price dynamics', Empirical Economics, 44(2), pp. 683-711.

Blake, A. P. and Mumtaz, H. (2017) 'Applied Bayesian econometrics for central bankers', Bank of England Centre for Central Banking Studies Technical Handbook No. 4. London: Bank of England. Available at: https://www.bankofengland.co.uk/ccbs/applied-bayesian-econometrics-for-central-bankers-updated-2017

Brignone, D. and Piffer, M. (2025) A structural VAR model for the UK economy. London: Bank of England, Macro Technical Paper No. 3. Available at: https://www.bankofengland.co.uk/macro-technical-paper/2025/a-structural-var-model-for-the-uk-economy

Carriero, A., Clark, T. E. and Marcellino, M. (2015) 'Bayesian VARs: Specification choices and forecast accuracy', Journal of Applied Econometrics, 30(1), pp. 46-73..

Carter, C. K. and Kohn, R. (1994) 'On Gibbs sampling for state space models', Biometrika, 81(3), pp. 541-553.

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

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998) 'Evaluating density forecasts with applications to financial risk management', International Economic Review, 39(4), pp. 863-883.

Domit, S., Monti, F. and Sokol, A. (2016) 'Forecasting the UK economy: Alternative forecasting methodologies and the role of off-model information', Bank of England Staff Working Paper No. 622. London: Bank of England.

Durbin, J. and Koopman, S. J. (2012) Time series analysis by state space methods. 2nd edn. Oxford: Oxford University Press.

Geweke, J. and Amisano, G. (2010) 'Comparing and evaluating Bayesian predictive distributions of asset returns', International Journal of Forecasting, 26(2), pp. 216-230.

Gneiting, T. and Katzfuss, M. (2014) 'Probabilistic forecasting', Annual Review of Statistics and Its Application, 1, pp. 125-151.

Gneiting, T. and Raftery, A. E. (2007) 'Strictly proper scoring rules, prediction, and estimation', Journal of the American Statistical Association, 102(477), pp. 359-378.

Harvey, A. C. (1989) Forecasting, structural time series models and the Kalman filter. Cambridge: Cambridge University Press.

Hubrich, K. (2005) 'Forecasting euro area inflation: Does aggregating forecasts by HICP component improve forecast accuracy?', International Journal of Forecasting, 21(1), pp. 119-136. d

Joseph, A., Potjagailo, G., Kalamara, E., Chakraborty, C. and Kapetanios, G. (2022) 'Forecasting UK inflation bottom up', Bank of England Staff Working Paper No. 915. London: Bank of England. Available at: https://www.bankofengland.co.uk/working-paper/2021/forecasting-uk-inflation-bottom-up

Kastner, G. (2016) 'Dealing with stochastic volatility in time series using the R package stochvol', Journal of Statistical Software, 69(5), pp. 1-30..

Kim, S., Shephard, N. and Chib, S. (1998) 'Stochastic volatility: Likelihood inference and comparison with ARCH models', Review of Economic Studies, 65(3), pp. 361-393.

Lütkepohl, H. (1987) Forecasting aggregated vector ARMA processes. Berlin: Springer-Verlag.

Nason, G. P. and Palasciano, H. A. (2025) ‘Forecasting UK consumer price inflation with RaGNAR: Random generalised network autoregressive processes’, International Journal of Forecasting (in press).

Omori, Y., Chib, S., Shephard, N. and Nakajima, J. (2007) 'Stochastic volatility with leverage: Fast and efficient likelihood inference', Journal of Econometrics, 140(2), pp. 425-449.

Ravn, M. O. and Uhlig, H. (2002) 'On adjusting the Hodrick-Prescott filter for the frequency of observations', Review of Economics and Statistics, 84(2), pp. 371-376.

Schwarz, G. (1978) 'Estimating the dimension of a model', Annals of Statistics, 6(2), pp. 461-464.

Stock, J. H. and Watson, M. W. (2007) 'Why has U.S. inflation become harder to forecast?', Journal of Money, Credit and Banking, 39(s1), pp. 3-33.

Stock, J. H. and Watson, M. W. (2010) 'Modeling inflation after the crisis', in Economic Policy Symposium Proceedings. Jackson Hole: Federal Reserve Bank of Kansas City, pp. 173-220.

Stock, J. H. and Watson, M. W. (2016) 'Core inflation and trend inflation', Review of Economics and Statistics, 98(4), pp. 770-784.

Stock, J. H. (2019). Trend, Seasonal, and Sectoral Inflation in the Euro Area (Working Paper). Princeton University.

Annex

Table 15. COICOP 2-digit CPI Divisions

Code	Division	Weights
01	Food and non-alcoholic beverages	113
02	Alcoholic beverages and tobacco	39
03	Clothing and footwear	60
04	Housing, water, electricity, gas and other fuels	128
05	Furnishings, household equipment and routine household maintenance	58
06	Health	28
07	Transport	132
08	Communication	24
09	Recreation and culture	149
10	Education	32
11	Restaurants and hotels	137
12	Miscellaneous goods and services	100

These weights show each category's share of average UK household spending (parts per thousand) in 2025. Recreation and culture has the highest weight (149), while communication have the lowest (24). Higher weights mean greater impact on overall inflation calculations.

Diebold-Mariano Test:

The Diebold-Mariano test (Diebold & Mariano, 1995) assesses whether two forecasting methods differ significantly in predictive accuracy. Suppose we have two models, m_1 and m_2 producing forecasts $\hat{Y}_{t,m}$ for the same observed series $\{Y_t\}_{t=1}^{T}$. We define each model's forecast errors as:

$$e_{t,m} = Y_t - \hat{Y}_{t,m}, \qquad m \in \{m_1, m_2\}$$

To measure forecast accuracy, a loss function is applied- commonly the squared error $L(e_t) = e_t^2$. At each time t, we then compute the loss differential:

$$d_t = {e^2}_{t,m_1} - {e^2}_{t,m_2}$$

The DM test evaluates whether the mean loss differential,

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t,$$

is significantly different from zero. Under the null hypothesis of equal predictive accuracy, $E(d_t) = 0$. If the p-value obtained from the DM test is below a chosen significance level (e.g., 5%), we reject this null, concluding that one model's forecasts are statistically more accurate. Otherwise, the models are considered indistinguishable in terms of forecast performance over the tested sample.

Density Forecast Evaluation

This annex sets out the technical definitions of the density forecast evaluation metrics used in Section 4.7. Let $F_{t,h}(\pi)$ denote the predictive cumulative distribution function (CDF) for inflation at forecast origin t and horizon h and let π_{t+h} denote the realised outcome. Posterior predictive draws are used to approximate these quantities.

Log Predictive Score (LS)

The log predictive score evaluates the probability density the model assigns to the realised outcome:

$$LS_{t,h} = \log f_{t,h}(\pi_{t+h})$$

Where $f_{t,h}$ is the predictive density function. Higher values indicate better forecasts, as more probability mass is placed at the outcome. The mean log score across the evaluation sample is reported.

Continuous Ranked Probability Score (CRPS)

The CRPS measures the squared distance between the predictive CDF and the realised outcome:

$$CRPS(F_{t,h}, \pi_{t+h}) = \int_{-\infty}^{\infty} (F_t(z) - \mathbf{1}\{\pi_{t+h} \leq z\})^2 dz$$

With posterior predictive draws $\{\pi_{t+h}^{(j)}\}_{j=1}^{N}$CRPS can be computed via:

$$CRPS(F_{t,h}, \pi_{t+h}) \approx \frac{1}{N}\sum_{j=1}^{N} \left|\pi_{t+h}^{(j)} - \pi_{t+h}\right| - \frac{1}{2N^2}\sum_{j=1}^{N}\sum_{k=1}^{N} \left|\pi_{t+h}^{(j)} - \pi_{t+h}^{(k)}\right|$$

Lower CRPS values correspond to sharper and better calibrated forecasts.

CRPS skill score

For model m relative to AR at horizon h:

$$Skill_h^{(m)} = 1 - \frac{\overline{CRPS}_h^{(m)}}{\overline{CRPS}_h^{(AR)}}$$

Where $\overline{CRPS}_h^{(m)}$is the mean CRPS across the forecast. A positive value indicates that model *m* outperforms the AR benchmark, while a negative value indicates worse performance. This skill score is scale-free, facilitating comparison across models and horizons.

Complementary dentistry accuracy tests

Coverage Tests

For a nominal coverage rate α (e.g. 50% or 90%), define the prediction interval $[L_{t,h}^{\alpha}, U_{t,h}^{\alpha}]$ Coverage is then defined as:

$$Coverage_{t,h}^{\alpha} = \mathbf{1}\{L_{t,h}^{\alpha} \leq U_{t,h}^{\alpha}\}$$

The empirical frequency of outcomes inside the interval is compared with the nominal coverage rate. Significant deviations indicate miscalibration.

Table 16. Coverage of 50% Predictive Intervals

Period	AR	UCSV-TD	UCSV-BU
Pre-COVID	36–64%	56–83%	56–72%
COVID / Energy	6–29%	30–46%	27–42%
Post-COVID	15–50%	64–88%	56–77%

Note**:** Coverage reports the proportion of realisations lying within the 50% nominal prediction intervals across horizons 1–6.

Probability Integral Transform (PIT)

The PIT statistic is defined as:

$$PIT_{t,h} = F_{t,h}(\pi_{t+h})$$

i.e. the cumulative probability assigned to the realised outcome. If predictive densities are correctly calibrated, PIT values are uniformly distributed on [0,1]. Departures from uniformity provide a diagnostic of over- or under-dispersion.

Table 17. PIT Diagnostics

Period	PIT Mean (AR)	PIT s.d. (AR)	PIT Mean (TD)	PIT s.d. (TD)	PIT Mean (BU)	PIT s.d. (BU)
Pre-COVID	0.54	0.33	0.57	0.19	0.61	0.21
COVID / Energy	0.54	0.33	0.57	0.19	0.61	0.21
Post-COVID	0.54	0.33	0.57	0.19	0.61	0.21

Note: A well-calibrated PIT should have mean ≈ 0.5 and standard deviation (s.d). ≈ 0.289 (uniform distribution). Higher standard deviation indicates over-dispersion, lower under-dispersion. Results are pooled across horizons.

UCSV Seasonality sensitivity test for harmonics.

Table 18. RMSE Comparison, Pre-COVID (2 vs 3 harmonics)

Horizon	AR	TD (2H)	TD (3H)	BU (2H)	BU (3H)
1	0.18	0.36	0.36	0.33	0.33
2	0.51	0.42	0.42	0.43	0.41
3	0.77	0.43	0.43	0.50	0.48
4	1.03	0.53	0.53	0.63	0.62
5	1.32	0.64	0.64	0.78	0.77
6	1.61	0.67	0.67	0.88	0.87

Notes: RMSEs of one-step to six-step-ahead forecasts. TD = top-down; BU = bottom-up. Two and three harmonics deliver almost identical results.

Table 19: Log Scores

Horizon	AR	UCSV-TD (2H)	UCSV-TD (3H)	UCSV-BU (2H)	UCSV-BU (3H)
1	0.238	-0.407	-0.407	-0.361	-0.318
2	-0.861	-0.723	-0.723	-0.637	-0.609
3	-1.552	-0.900	-0.900	-0.845	-0.798
4	-2.063	-1.049	-1.049	-1.096	-1.025
5	-2.388	-1.183	-1.183	-1.313	-1.257
6	-3.004	-1.262	-1.262	-1.545	-1.489

Sensitivity Analysis: Prior Specification

Table 20. Forecast Differences vs Baseline

(a) Tight vs Baseline

Horizon	Mean Diff	RMSE Diff	Max Abs Diff
1	0.02	0.06	0.14
2	-0.01	0.07	0.16
3	0.03	0.08	0.22
4	0.00	0.08	0.14
5	0.01	0.05	0.12
6	0.00	0.05	0.12

(b) Loose vs Baseline

Horizon	Mean Diff	RMSE Diff	Max Abs Diff
1	-0.03	0.10	0.30
2	-0.05	0.10	0.27
3	-0.02	0.08	0.17
4	-0.04	0.07	0.16
5	-0.02	0.08	0.15
6	-0.03	0.09	0.19

ANNEX B: KSB MAPPING

Plagiarism Statement: I confirm that this report is my own work, is not copied from any other person's work (published or unpublished) and has not previously been submitted for assessment anywhere. I have read and understood the EPAO's regulations on plagiarism.

High Level KSBs	Individual KSBs	Description
Applied economic analysis	**S1** Apply micro-economic and macro-economic theories and modelling, including econometric, to inform a range of business and policy decisions.	The 'Methodology' section details the econometric framework used, specifically Bayesian unobserved components models with stochastic volatility (UCSV) to decompose inflation into trend, seasonal, and irregular components. The econometric results in Section 5 directly inform policy recommendations regarding short-term inflation forecasting, specifically addressing Bank of England forecasting challenges identified in the Bernanke (2024) review. The theoretical foundation draws on macroeconomic time series analysis, particularly the literature on inflation persistence and volatility modelling (Stock & Watson, 2007; 2016, 2019).
Applied economic analysis	**S2** Convert the policy or other question into a tractable appraisal, evaluation or other analysis drawing on the most appropriate analytical method, including non-market valuation methods. Analysis considers, inter alia: the counterfactual, opportunity cost, risk and uncertainty and how to estimate discount rates and costs of capital.	I underwent an iterative methodology development process with my dissertation supervisor, seeking his steers on questions I had. This demonstrates systematic conversion of the complex policy question of inflation forecasting improvement into a tractable analytical framework, weighing trade-offs between different estimation approaches. The decision to develop a custom Bayesian UCSV implementation rather than use existing packages (statsmodels UC, pybuc, R bsts) was driven by critical analytical requirements that off-the-shelf solutions could not meet. Specifically, the project's emphasis on density forecast evaluation required full Bayesian estimation to generate proper predictive distributions, while the stochastic volatility capability essential for capturing time-varying uncertainty was unavailable in the Python alternatives. This methodological choice involved significant development costs but was necessary to deliver policy-relevant uncertainty quantification that existing tools could not provide, demonstrating appropriate evaluation of analytical trade-offs against resource constraints The recursive pseudo-real-time evaluation design explicitly addresses the counterfactual of what forecasters would have known at each point in time, ensuring realistic assessment of model performance in policy-relevant conditions. Risk and uncertainty analysis is central to the approach through comprehensive density forecast evaluation, reflecting the policy importance of probabilistic guidance rather than point forecasts alone, directly

		addressing Bernanke review recommendations for improved uncertainty communication.
	S3 Critically assess available information sources and judge validity and usefulness for the issue at hand; clean and manipulate data; be aware of data limitations and explain them; clearly describe and present data using data visualisation techniques; and draw out and explain policy and business implications to clients.	The Literature Review critically evaluates existing inflation forecasting approaches, from simple autoregressive models to machine learning methods, determining which elements are most relevant for UK policy context. The 'Data' section provides detailed assessment of ONS CPI data sources, explaining the choice of non-seasonally adjusted monthly data and the treatment of basket weights for bottom-up aggregation. Figures 1-3 demonstrate effective data visualisation, showing inflation dynamics across different regimes and seasonal patterns by CPI division. The 'Discussion' section clearly draws out policy implications, specifically addressing the Bank of England's forecasting toolkit and the complementary role of UCSV models alongside expert judgement.
Project management and planning	**S5** Scope areas of work identifying: objectives, analytical methods, resources required and potential delivery risks. Able to recognise when complementary expertise is required e.g. scientists, other social scientists and data specialists.	Project scope was systematically developed through targeted stakeholder engagement within HMT's GDP & Inflation branch: Head of Inflation consultation clarified policy needs, while Head of GDP provided technical guidance on my choice of models. Resource constraints required careful planning: custom R development demanded significant time investment but was necessary when existing packages couldn't meet stochastic volatility requirements.. Multi-disciplinary expertise was strategically sought: Professor Guy Nason (Imperial College London, co-author of machine learning UK CPI forecasting research) provided insights on alternatives models and methodological trends, while the head of my team (a badged statistician) expressed an interest in the state space model. My work organisation was structured using systematic project management tools: a Kanban board provided workflow management with columns for "To Do," "In Progress," and "Completed" tasks, enabling clear tracking of research phases from literature review through robustness testing. The MoSCoW prioritisation framework (Must have, Should have, Could have, Won't have) was applied to manage scope creep and resource allocation - identifying density forecast evaluation and recursive real-time assessment as "Must have" requirements, seasonal specification robustness as "Should have," while additional harmonics testing and additional model comparison were classified as "Could have" enhancements. Potential risks are identified in the 'Discussion', including the limitations of univariate models for

		scenario analysis and the continued importance of expert judgement for incorporating policy changes. .
Effective communication	**S7** Clearly communicate economic principles and concepts to non-economists; present trade-offs and uncertainties and articulate these clearly; frame advice, drawing on knowledge of stakeholders' positions, for maximum impact.	The 'Non-technical Summary' translates complex econometric concepts into accessible language for policy audiences, avoiding jargon while maintaining analytical precision. Trade-offs between different approaches are clearly presented in the comparison of top-down versus bottom-up forecasting, explaining why aggregation can sometimes amplify rather than reduce forecast errors. Uncertainty quantification was prioritised, with comprehensive density forecast evaluation addressing policymakers need for probabilistic guidance during volatile periods rather than misleading point estimates that had failed during the energy price shock for example. The analysis is framed with explicit awareness of central bank needs (such as the Bank of England) stakeholder needs, addressing recommendations from the Bernanke review and positioning UCSV as a complement to existing tools rather than a replacement.
Horizon scanning	**S6** Use horizon scanning methodologies to anticipate new trends, opportunities and challenges that may influence outcomes of interest to client.	The 'Conclusion' identifies future research directions, including the potential for incorporating explicit outlier detection and testing forecast combination approaches as the forecasting environment continues to evolve. Discussion of machine learning approaches and their emerging role in central bank forecasting reflects awareness of technological developments that may influence future practice.
Maintaining quality standards	**S8** Design Quality Assurance processes and implement these, following organisational best practices, and drawing on sources of external expertise; critically assess economic analysis and improve it.	Comprehensive robustness checks are implemented, including sensitivity analysis for seasonal specification (2 vs 3+ harmonics), prior specification sensitivity, and HP filter initialisation parameters. Multiple evaluation metrics are employed (RMSE, log scores, CRPS, coverage tests, PIT diagnostics) following best practice in forecast evaluation literature to ensure robust assessment. Statistical significance testing via Diebold-Mariano tests provides rigorous comparison against benchmarks, following established econometric standards. The recursive pseudo-real-time evaluation design follows central banking best practices for forecast assessment.
	B1 Ethical conduct: analyst attributes sources and ideas to their originator; provides honest advice on all	All sources are properly cited with comprehensive bibliography acknowledging the foundational work of Stock & Watson and other key contributors to unobserved components literature.

	relevant aspects to an issue; avoids bias.	Limitations are honestly acknowledged, including the poor one-month horizon performance and the constraints of univariate modelling for policy scenario analysis. Bias is avoided by using an established autoregressive benchmark and transparent evaluation criteria, with results reported even when they contradict expectations (e.g., bottom-up deterioration at longer horizons).
	B4 Rigour: demonstrates a commitment to detail.	Extensive robustness testing across multiple dimensions: seasonal specification, prior sensitivity, and initialisation choices, with results reported in annexes to maintain transparency while preserving readability. The analysis covers three distinct regimes to test model stability across different macroeconomic environments, ensuring findings are not driven by sample-specific features. Both point and density forecast evaluation provides comprehensive assessment, recognising that policy users need both central projections and uncertainty measures. Technical implementation details are provided in methodology section, including specific priors and MCMC procedures, enabling replication and validation of results.

School of Economics and Finance

Queen Mary University of London

This working paper is based on project work undertaken by EMAP apprentices

School of Economics and Finance
Queen Mary University of London
Mile End Road
London E1 4NS
Tel: +44 (0)20 7882 7356
Fax: +44 (0)20 8983 3580
Web: www.econ.qmul.ac.uk/research/workingpapers/