Bank of England

Learning from forecast errors: the Bank's enhanced approach to forecast evaluation

Macro Technical Paper No. 6
January 2026

Raphael Abiry, James Hurley, Paul Labonne, David Latto, Harry Li, Andre Moreira, Joseph Oyegoke and Sumer Singh

A series designed to document models, analysis and conceptual frameworks for monetary policy preparation – they are written by Bank staff to encourage feedback and foster continued model development.

Macro Technical Paper Series

Dr Bernanke's 2024 **Review** of the monetary policymaking processes at the Bank of England provided a number of constructive recommendations for reform which we are taking forward.

This included improving our model maintenance and development. Macroeconomic models and frameworks play an important role in the monetary policy process. To maximise the value of macroeconomic models, they must be well documented and continuously improved as time goes on.

This Macro Technical Paper (MTP) series is part of the Bank's response to Dr Bernanke's recommendations. These MTPs are authored by Bank staff, and are intended to document models, analysis, and conceptual frameworks that underpin monetary policy preparation. The models documented in the series will typically be used to assess the current state of the economy, forecast its future, and to simulate alternative paths and policy responses.

Importantly, while each MTP will provide insights about a particular model or modelling framework that is an 'input' to policy, no single model can possibly capture all the relevant features to perform even just one of those roles adequately. Models will inevitably have to be updated and will improve over time, including as they are adapted to different constellations of macroeconomic conditions. This is a natural part of real-time monetary policy making. So, inevitably, no MTP will provide definitive answers.

The Bank seeks to encourage an active and informed debate about its modelling frameworks. Publishing and discussing the analytical work undertaken to support its monetary policy choices is central to this ambition. These MTPs will support a culture of continuous learning in monetary policy making. As time goes on, Bank staff will update and upgrade models, drawing on insights from the frontier of the academic literature. Moreover, this transparency will encourage external engagement from experts to ensure our modelling tools remain fit for purpose.

Clare Lombardelli
Deputy Governor for Monetary Policy

Bank of England

Macro Technical Paper No. 6

Learning from forecast errors: the Bank's enhanced approach to forecast evaluation

Raphael Abiry,[1] James Hurley,[2] Paul Labonne,[3] David Latto,[4] Harry Li,[5] Andre Moreira,[6] Joseph Oyegoke[7] and Sumer Singh[8]

Abstract

The Bernanke Review and forecasting challenges of recent years have highlighted the importance of continuous learning from forecast errors. Following substantial investment by Bank staff, and alongside the publication of a new Forecast Evaluation Report in 2026, this paper provides technical detail on the Bank's enhanced forecast evaluation approach. We describe a wide range of evaluation techniques allowing us to characterise the Bank's forecast performance statistically, as well as to interrogate specific forecast errors in greater detail, including their economic drivers. Worked examples are provided throughout, focusing on a subset of macroeconomic variables relevant to monetary policy makers. The statistical techniques described in this paper are also implemented and published as part of a new Python package, to facilitate ongoing forecast evaluation and continued development of this toolkit.

Key words: Forecast evaluation, macroeconomic forecasting, forecast accuracy, model validation, central banking, monetary policy.

JEL classification: C52, C53, C54, C55, E17.

---

(1) Bank of England. Email: raphael.abiry@bankofengland.co.uk
(2) Bank of England. Email: james.hurley@bankofengland.co.uk
(3) Bank of England. Email: paul.labonne@bankofengland.co.uk
(4) Bank of England. Email: david.latto@bankofengland.co.uk
(5) Bank of England. Email: harry.li@bankofengland.co.uk
(6) Bank of England. Email: andre.moreira@bankofengland.co.uk
(7) Bank of England. Email: joseph.oyegoke@bankofengland.co.uk
(8) Bank of England. Email: sumer.singh@bankofengland.co.uk

We thank colleagues at the Bank of England for their careful review of this paper and helpful comments, including Abigail Haddow, Alessandro Morico, Huw Pill, Kate Reinold and Tim Willems. We also thank Lennart Brandt, Clara Churchill and Diego Lopez for their contributions to refining the newly developed forecast evaluation toolkit. Additionally, we thank Jennifer Castle, Jack Fosten and Ana Galvão for their technical guidance on the evaluation techniques and approaches employed here. The views expressed in this paper are those of the authors and do not necessarily represent those of the Bank of England or any of its committees. The techniques described in this paper have been implemented in an open-source Python package available at https://github.com/bank-of-england/forecast_evaluation.

The Bank's macro technical paper series can be found at
www.bankofengland.co.uk/macro-technical-paper/macro-technical-papers

Bank of England, Threadneedle Street, London, EC2R 8AH
Email: enquiries@bankofengland.co.uk

©2026 Bank of England
ISSN 2978-3194 (online)

Contents

1 Introduction

The Bank of England has published a quarterly set of economic forecasts since the introduction of inflation targeting in 1992. Since 1997, when the Bank gained operational independence to set monetary policy, these forecasts have been used as an input to the Monetary Policy Committee's (MPC) decisions. Economic forecasts remain an important input to the MPC's deliberations, although their role in the policymaking process is evolving as part of the Bank's response to the recent review of 'forecasting for monetary policy making and communication at the Bank of England', led by Dr Ben Bernanke (Bernanke, 2024; Dhami et al., 2025).

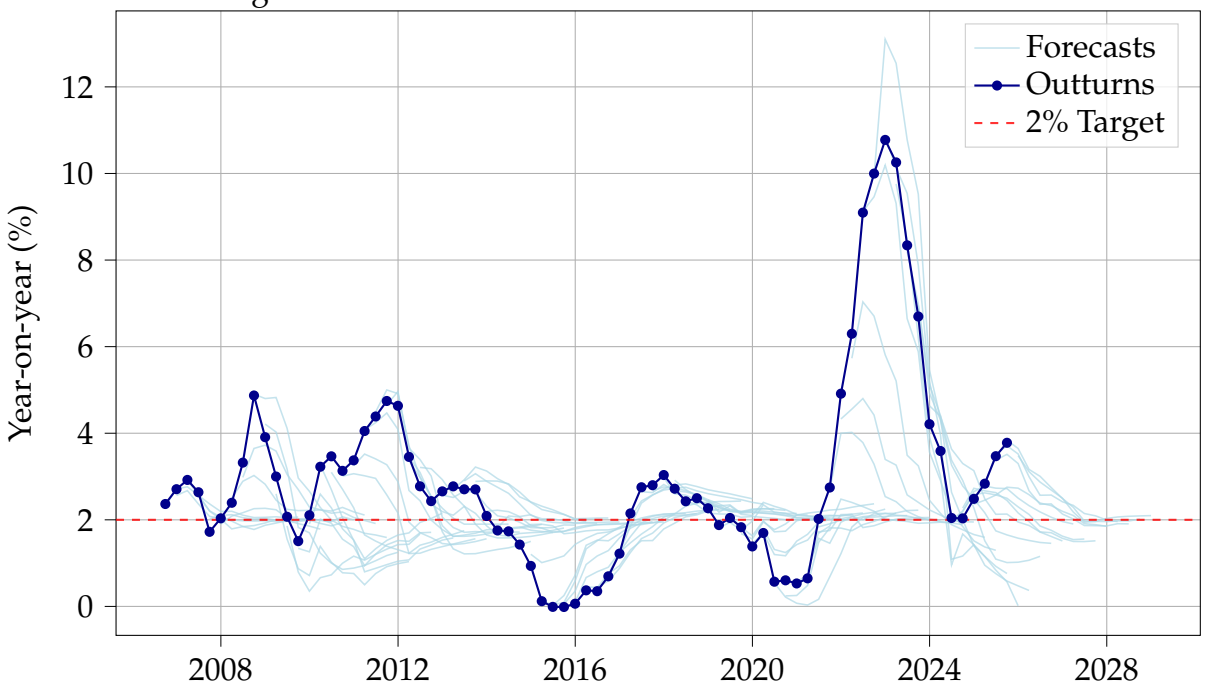
Even as forecasts are increasingly supplemented by a range of other analytical inputs such as scenarios, evaluating how those forecasts perform can be an important driver of 'continuous learning' (Lombardelli, 2024). By providing rigorous and regular feedback on the Bank's forecast performance, systematic forecast evaluation can help guard against persistent biases or errors, and be leveraged more generally to drive improvements to our models, processes, and understanding of the economy. In doing so, it can ultimately also help support better informed monetary policy decisions.

The Bank has a long history of forecast evaluation. For more than a decade, Monetary Policy Reports had featured an annual review of recent forecast errors (for latest examples see Bank of England, 2023, 2024). Less frequently, evaluations of longer-term forecast performance have also been carried out, both officially by the Bank's Independent Evaluation Office (Independent Evaluation Office, 2015), and more recently in staff research (Kanngiesser and Willems, 2024). Fuller reviews of the Bank's forecasting capability have at times been commissioned from external experts to draw lessons from significant historical experiences such as the global financial crisis or the aftermath of the Covid-19 pandemic (Stockton, 2012; Bernanke, 2024).

The Bernanke review focused specifically on the Bank's forecasting during the post-Covid inflation episode. This was a particularly challenging time for economic forecasters and the Bank of England was no exception, as our central projections significantly underestimated the scale and persistence of the upswing in inflation that unfolded over 2021 and 2022 (Figure 1). While the review generally found the Bank had performed no worse than other UK forecasters or peer central banks, it nevertheless identified a number of potential avenues for improvement. Among these was a clear recommendation to strengthen our forecast evaluation practices.

> Recommendation 5: The staff should be charged with highlighting significant forecast errors and their sources, particularly errors that are not due to unanticipated shocks to the standard conditioning variables. Models and model components that may have contributed to forecast misses should be regularly evaluated and discussed, as well as the determinants of variables whose forecasts are consistently dominated by extra-model judgements. Staff should routinely meet with MPC members to consider whether structural change, misspecification of models, or faulty judgements warrant discrete changes to the key assumptions or modelling approaches used in forecasting. — Bernanke (2024)

Figure 1: MPR forecasts and outturns of CPI inflation



The light blue lines show forecasts from different vintages. The dark blue line shows the outturns $k = 12$ quarters after the first data release.

As the quote suggests, one of the keys to harnessing the benefits of forecast evaluation is the ability to distinguish between different sources of forecast error. In that context, it is important to recognise that some forecast errors are unavoidable, particularly those that are due to shocks occurring after forecasts were made. As an illustration of the role that such unanticipated shocks can play, Bank staff estimate that roughly half of the inflation undershoot in forecasts made in late 2021 can be attributed to energy price rises linked to Russia's invasion of Ukraine in early 2022 (Bank of England, 2026). At the same time, there are several other error sources that forecast evaluation can help to identify and address. These include model misspecification of various kinds, flaws in the formulation of 'conditioning paths' fed exogenously into models, or miscalibration in judgements that are sometimes also overlaid to adjust for off-model information or known model limitations.

In practice, drawing the right lessons from forecast evaluation is not straightforward and requires using a range of complementary techniques. With different approaches better suited to different questions, and each with their own uncertainties and assumptions, no single approach is likely to provide definitive answers. Instead, users must often weigh several competing signals and exercise careful judgement in forming conclusions. This Macro Technical Paper describes a wide array of techniques that Bank staff have implemented to support ongoing forecast evaluation and underpin the production of the Bank's new Forecast Evaluation Report (FER, Bank of England (2026)).

Building on past work from Independent Evaluation Office (2015) and Kanngiesser and

Willems (2024), a first set of techniques allows us to characterise the Bank's forecast performance statistically over longer evaluation windows. These techniques can help to detect ingrained forecast deficiencies, as well as place more recent forecast errors into context. To do this, we rely on a suite of accuracy, unbiasedness, and efficiency tests that are standard in the literature (Mincer and Zarnowitz, 1969; Nordhaus, 1987; Diebold and Mariano, 1995). To control for the effects of economic shocks that can sometimes make forecasting inherently more difficult, such as the energy price rises of 2022, we also enable systematic comparisons to a rich set of benchmark models. These include an AR(p) approach, a Bayesian VAR, and a purely mechanical implementation of the Bank's recently improved DSGE model known as COMPASS (Albuquerque et al., 2025), all of which are vulnerable to those shocks in the same way.

We then discuss a range of more targeted approaches enabling us to interrogate specific forecast errors and their drivers, including the role of conditioning path news. Identifying and addressing forecast deficiencies in as close as possible to real time is essential for maximising the value of forecast evaluation to policymakers, but traditional forecast evaluation approaches require longer samples to provide meaningful conclusions. A more flexible toolkit is therefore needed to supplement real-time forecast error analysis. Here we do not seek to be exhaustive, but illustrate these ideas using a subset of techniques that Bank staff have used recently, including in the production of the 2026 Forecast Evaluation Report. These include distributional analysis of errors, rolling-window 'fluctuation' tests, analysis of the role of data revisions, counterfactual analysis of the role of conditioning paths, and other model-based decompositions.

Uses in practice

As part of the ongoing response to the Bernanke Review and its wider Monetary Policy Transformation programme (Lombardelli, 2024), the Bank of England is publishing a new Forecast Evaluation Report (FER), for which this Macro Technical Paper serves as the technical companion. Compared with past Bank forecast evaluations, the FER has a significantly expanded scope, combining a comprehensive assessment of the Bank's longer-run forecast performance with more detailed analysis of recent misses.

At the same time, Bank staff are actively exploring how to incorporate more frequent insights from forecast evaluation into the MPC's quarterly policy round. The aim is to prompt real-time improvements to modelling and forecasts, inform ongoing discussions of forecast uncertainties and judgements, and where appropriate also motivate other relevant analytical work, including scenario analysis.

To support this, Bank staff have developed a new forecast evaluation toolkit, integrating a range of established forecast evaluation methods with best practices from data science to enable the efficient handling of large volumes of data and forecast vintages (the baseline dataset attached to the toolkit contains about two million records). This allows us to evaluate and benchmark the Bank's forecasts against a range of model-based alternatives on the basis of real-time information. The toolkit implements the full set of statistical evaluation techniques described in this Macro Technical Paper, but not the more targeted economic approaches.

Bank staff will continue to develop this toolkit in light of experience, changes in the economic environment, and advances in the forecast evaluation literature. Its modular design and extensive suite of unit tests makes it easily extendable to additional techniques. It has also been designed to be applicable to a wider range of forecasting contexts, supporting more standardised evaluation in model development and communication of forecast results. Finally, the toolkit is being released as an open-source Python package. We hope it will be useful to researchers and practitioners evaluating their own models, and welcome feedback and contributions from the wider community.

Outline

The remainder of the paper is organised as follows. Section 2 reviews relevant literature on applied forecast evaluation and places the Bank's refreshed approach in its context. Section 3 describes relevant aspects of the Bank's forecasts alongside a range of benchmark models used to support their evaluation. Section 4 describes a range of statistical techniques utilised to assess long-term historical forecast performance. Section 5 discusses a few more targeted approaches that can be used to interrogate more recent forecast errors. Section 6 concludes.

2 Selected literature: evaluation approaches, findings and new directions

The academic literature on forecast evaluation offers a wide set of statistical tests and metrics of forecast performance (Diebold and Mariano, 1995; Mincer and Zarnowitz, 1969; Blanchard and Leigh, 2013; Nordhaus, 1987). Historical forecast performance is typically evaluated along three dimensions. First, accuracy, which measures how close forecasts are to realised economic outcomes, often relative to a set of competitors. Second, unbiasedness, which assesses whether forecasts systematically over- or under-predict outcomes. Third, efficiency, which examines whether forecasters utilise available information optimally when forming and revising predictions. These statistical metrics can collectively provide a helpful assessment of forecast quality to identify areas for potential improvement in the forecasting process.

Our new Python-based forecast evaluation toolkit (appendix A) implements some of the most widely used statistical approaches from the literature, described in detail in section 4, and applies them to a large dataset of forecast errors. We also utilise a range of more targeted approaches, including other model-based insights to investigate the economic drivers of forecast errors, as discussed in section 5. To help place the Bank's enhanced forecast evaluation approach in context, the remainder of this section discusses a range of past evaluations of our Monetary Policy Report (MPR) forecasts, both internal and external, as well as forecast evaluations carried out at other UK institutions and international central banks. We highlight the key contributions we build upon and some promising avenues for future development.

2.1 The Bernanke review and other past evaluations of the Bank's MPR forecasts

Large shocks in the post-Covid period, particularly to energy prices, made forecasting especially challenging. This formed the backdrop for the Bernanke review (Bernanke, 2024), which provided recommendations for how the Bank's forecasting processes and systems could be revamped to better support the Monetary Policy Committee's decision making, especially in times of high uncertainty.

As part of the review, Bernanke (2024) compared the forecast performance of the Bank of England with other major central banks, and a range of external forecasters, focussing on the accuracy of projections for inflation, GDP growth and the level of the unemployment rate. He found that the Bank's forecast performance had deteriorated with the onset of the pandemic and the subsequent inflation. However, this deterioration in forecast accuracy since the pandemic had been similar for external forecasters and other major central banks. This suggests that the deterioration in forecast performance in the post-Covid period can generally be attributed to heightened economic volatility, rather than particular deficiencies in the Bank's forecasting framework.

There have been several other notable evaluations of the Bank's MPR forecasts over the years, including internal reports and external evaluations within the broader academic literature. We summarise the key results and our learnings from these evaluations in further detail below.

2.1.1 Internal evaluations of MPR forecasts

Independent Evaluation Office (2015) utilised a similar set of statistical evaluation techniques to those implemented in our new Python toolkit (these were further independently reviewed at the time by Professor James Mitchell at the University of Warwick). The IEO report specifically explored accuracy, bias and efficiency in MPR forecasts of key macroeconomic variables, including GDP growth, CPI inflation, wage growth and the unemployment rate up to 2014. It found the Bank's forecasts for GDP and inflation had been more accurate than private-sector and other central-banks forecasts at most horizons. It did not find evidence of bias in the Bank's forecasts for those variables, or of inefficiencies up to one-year ahead. By contrast, wage growth and unemployment forecasts were found to be somewhat biased, with unemployment forecasts performing worse than both the private sector and other central banks. Wage growth and unemployment forecasts were also found to be inefficient.

Staff research by Kanngiesser and Willems (2024) extended parts of the IEO's analysis to a more recent sample of forecasts ending in 2024. This research benchmarked the Bank's MPR forecasts against two simple statistical models – a random walk and a univariate autoregressive model – and found the MPR to be relatively accurate in comparison. However, it also found some evidence of bias in the Bank's forecasts for GDP growth, wage growth and unemployment. Kanngiesser and Willems (2024) further proposed using Blanchard and Leigh (2013) regressions to investigate whether or not some key economic relationships may have miscalibrated on average within the Bank's forecasts. Results suggested the Bank's forecasts may have underestimated the pass-through

of real wage growth and unemployment to CPI inflation, and the speed of monetary policy transmission. We follow Kanngiesser and Willems (2024) in comparing the MPR with statistical benchmarks and implementing Blanchard and Leigh (2013) regressions to explore cross-variable inefficiencies.

Before launching its more comprehensive Forecast Evaluation Report in 2026 (Bank of England, 2026), the Bank had also published an annual (with some notable exceptions, for example, during Covid-19) analysis of recent forecast errors as part of its Monetary Policy Report, and formerly Inflation Report, publications (for latest examples see Bank of England, 2023, 2024). These exercises focused primarily on forecast errors made over the past year, exploring some of the reasons for those misses, including the role of conditioning paths. We explore and expand on some similar techniques in this paper and in the 2026 FER.

2.1.2 External evaluation of the Bank's forecasts in the academic literature

The Bank's historical forecasts have also been evaluated extensively in the academic literature (Clements, 2004; Wallis, 2004; Boero et al., 2008; Groen et al., 2009). This section summarises some more recent papers.

Coroneo (2025) focused on analysing the accuracy of the Bank of England's inflation forecasts relative to model benchmarks and external forecasters. Compared to our current forecast evaluation toolkit, she exploited a richer set of error metrics, notably accounting for asymmetric preferences by penalising some kinds of errors more than others. She found that the Bank's forecasts consistently outperformed model benchmarks but did not systematically beat external forecasters, except for one-year-ahead forecasts where external forecasters had tended to over-predict inflation significantly compared to the Bank. In addition to comprehensive error metrics, Coroneo (2025) suggested implementing fluctuation tests to investigate local variations in relative performance, which we also implement in our new toolkit.

Castle et al. (2025) introduced a set of techniques that allow for real-time detection of shifts in economic trends using successive forecast errors, which can then be used to adjust future forecasts. Applying these techniques to the Bank's inflation forecasts in the post-pandemic period, they found that forecasts based on real-time detection of trend breaks could have been more accurate than the Bank's MPR forecasts over the same period. This could be a valuable extension to our toolkit in future.

Looking beyond the UK, Argiri et al. (2024) conducted a broader evaluation of the Bank of England's, European Central Bank's and US Federal Reserve's forecasts of inflation since 2000. The key finding from this comparison were that all three central banks' inflation forecasts had been relatively accurate compared to external forecasters, unbiased and efficient, particularly at shorter time horizons.

2.2 Forecast evaluation at other UK institutions and international central banks

The UK's Office for Budget Responsibility (OBR) has published a comprehensive annual Forecast Evaluation Report (FER) since 2011. The OBR's approach to evaluating forecast accuracy has relied primarily on comparisons with other forecasters, including the Bank of England. The OBR's FERs also reflect on the reasons for divergence between forecasts and outturns, drawing lessons from forecast errors and broader challenges to inform model development priorities. For example, lessons from their 2024 FER (see Office for Budget Responsibility, 2024) include refining the inflation forecast to better capture indirect energy price effects and developing a suite of wage equations to help forecast labour market developments. We plan to follow the OBR's lead on this, building stronger feedback mechanisms from evaluation to modelling priorities.

Like the Bank of England, other central banks have also published evaluations of their forecasts following the recent spike in inflation across advanced economies.

Bohm and Sing (2022) conducted an evaluation exercise for the Reserve Bank of New Zealand (RBNZ), focusing on the Covid-19 period. In particular, they compared the RBNZ's forecasts to a range of large private sector institutions in New Zealand, finding that both sets of forecasts had been similarly accurate over this period of significant economic disruption. Johansson et al. (2023) also evaluated the Riksbank's macroeconomic forecasts against private sector institutions, focusing on performance during the post-pandemic period of 2021-22. They likewise found that Riksbank forecasts over this period had been as or more accurate than private sector forecasts.

In Norway, Bowe et al. (2023) have specifically evaluated the performance of Norges Bank's SMART forecasting toolkit. This was compared against standard statistical benchmarks, and focussing on the accuracy of inflation and GDP growth forecasts. SMART is an integrated toolkit where forecasting and evaluation are performed seamlessly in real time, utilising and combining forecasts from a very large number of models. They found forecasts from this system to have similar or better performance than the benchmarks they use. Our forecast evaluation toolkit takes some inspiration from SMART's flexible and modular approach, allowing real time applications and easy addition of new benchmark models and metrics in future. We are currently also exploring ways in which we can integrate this toolkit more fully with our broader forecasting infrastructure.

In addition to statistical evaluations, the European Central Bank (ECB) has also used a range of model-based economic approaches to interrogate the drivers of large forecast errors made during the recent inflation surge, including the role of exogenous surprises to key conditioning paths (Lane, 2024). They have further used this approach to analyse the potential role of data uncertainty and endogenous features of forecasting models. We describe a similar approach to investigating the role of conditioning assumptions in section 5 of this paper.

2.3 The Bank's new FER and forecast evaluation toolkit in context

The Bank's new Forecast Evaluation Report (Bank of England, 2026), launched alongside this Macro Technical Paper, marks a significant step forward for our forecast evaluation practices, responding to an important recommendation of the Bernanke review. It builds on previous Bank forecast evaluations (Independent Evaluation Office (2015) and Kanngiesser and Willems (2024)) by combining insights from longer-term historical evaluations with more targeted analysis of recent errors.

The Python-based forecast evaluation toolkit that Bank staff have developed to underpin this is also designed to help embed regular forecast evaluation into the Bank's internal processes. The toolkit brings together data science methods and established techniques from the economics literature to evaluate economic forecasts. It also draws inspiration from the practice of other institutions, and from our own past experience of internal evaluation. But we have sought to innovate along three key dimensions.

First, by bringing together a vast set of evaluation metrics and techniques under a single integrated toolkit, making it widely available for both internal and external use. The toolkit enables the efficient handling of large volumes of relevant data and forecast vintages, allowing forecasts to be evaluated using relevant real-time information. This includes benchmarking the Bank's forecasts against model-based alternatives that are constructed on the same real-time basis, enabling meaningful comparisons.

Second, by complementing those metrics with an unusually rich set of benchmark models. These include not only standard 'naive' benchmarks, but also real-time backtested forecasts from a Bayesian VAR model that is used regularly in cross-checking the Bank's baseline forecasts, and from our workhorse DSGE model known as COMPASS (Albuquerque et al., 2025). The inclusion of a hands-free version of COMPASS in particular should enable us to consider the joint role of judgements and other aspects of the Bank's forecast process that take place outside of COMPASS, for example in a range of other suite models (Burgess et al., 2013).

Finally, through a fruitful collaboration between economists and data scientists, we have sought to leverage modern data engineering techniques to implement this new toolkit in a way that is robust, scalable, and accessible. A pilot version of a dashboard drawing directly on this toolkit has already been developed, providing an easy-to-use front end for both staff and the MPC to explore insights from forecast evaluation. This architecture also sets us up more generally to continue developing the Bank's forecast evaluation toolkit in the future, for example by integrating further tests or benchmark models.

3 MPR forecasts and benchmarks

The Bank's forecasts project key macroeconomic variables up to twelve quarters ahead, using a range of models, conditioning assumptions, and expert judgement. They are published quarterly in the Monetary Policy Report (MPR). Section 3.1 discusses some relevant features of the forecast. Section 3.2 describes our selection of benchmark

models. Section 3.3 discusses real-time data and estimation settings, explaining how we take the models to the data to be as consistent as possible with the MPR forecasts.

3.1 The MPR forecast

This section summarises some key features of MPR forecasts with implications for its evaluation. For more information, see section 2 of the Forecast Evaluation Report (Bank of England, 2026).

3.1.1 What the MPR forecasts represent

Point forecasts published by the Bank in recent years have corresponded to the 'modal', or single most likely path for the economy. Note that modal forecasts differ from the mean, for example if there is a skew in the forecast distribution, but in practice – with some notable exceptions – differences tend to be small. Historically, Bank forecasts were conceptually closer to the mean, derived from model outputs without explicit probabilistic framing. The shift to the mode came in the early 2000s, as part of a broader effort to improve transparency and better communicate uncertainty. While mean projections are still produced and published in supplementary materials, they are usually not the primary forecast communicated to the public. Bernanke (2024) also highlighted the importance of clarity around what the forecasts represent, and a rigorous process for introducing differences between mean and modal forecasts.

The forecasts have tended to represent the best collective judgement of the MPC since relatively early in the foundation of the committee (Bank of England, 1999). Agreement on the forecast was achieved through a thorough process of discussion and layering of judgements. While the broad spirit of that process remains in place today, the role of the forecast within the policymaking process is changing once again, with recent Bernanke Review reforms moving that towards being one among 'multiple inputs' to policy, and the published forecast (based on a staff proposal) containing projections that a majority of the MPC agree are a reasonable baseline (Bailey, 2025). The emphasis on best collective judgement has been reduced as a result (see Dhami et al., 2025; Haberis et al., 2025; Alati et al., 2025). As noted in the foreword to the Forecast Evaluation Report (Bank of England, 2026), from a policymaking perspective, the forecast should not just perform well in statistical tests, its key purpose is to support the MPC in formulating and communicating policy decisions internally and externally.

It is important to note that the Bank's forecasts are conditional on a number of assumptions that are detailed further in section 3.1.3. One implication of this approach is that MPR forecasts may not always correspond to the MPC's single most likely expectation, particularly where conditioning paths for one or more variables may not correspond to the MPC's own views of their most likely evolution. This can also result in temporary inconsistencies between conditioning paths, which can in turn lead to predictable forecast errors. For example, if fiscal policy was widely anticipated to loosen relative to the existing conditioning assumptions, financial market participants might expect a higher Bank Rate path, all else equal. As the Bank's forecasts condition on these financial market expectations for Bank Rate, that might show up as a drag on the forecast for

GDP growth without the role of potentially looser fiscal policy being fully incorporated into the forecast.

3.1.2 The role of models

The Bank's New Keynesian (NK) Dynamic Stochastic General Equilibrium (DSGE) model known as 'COMPASS' has long sat at the heart of producing the MPR forecasts. However, this process also draws regularly on a much broader array of other models. Collectively, these models provide a helpful starting point in processing the latest developments in the domestic and global economic conjunctures, before being supplemented by staff and policymaker judgement.

At short horizons, the Bank's forecasts are informed by statistical models, which improve forecast performance by incorporating available high-frequency information. Staff deploy a range of nowcasting techniques to produce these short-term forecasts. Past Bank research has documented our use of bridge models, dynamic factor models and mixed data sampling – or MIDAS – models to nowcast GDP (Bell et al., 2014; Anesti et al., 2017). More recently, a more bespoke Staggered-Combination MIDAS approach, designed to optimally exploit the properties of UK official and survey data, has been used to inform GDP and labour market nowcasts (Moreira, 2025; Daniell and Moreira, 2023). Short-term inflation forecasts rely heavily on a set of Unobserved-Component models, dealing with particular challenges of UK CPI data such as seasonality, time-varying trends, and stochastic volatility (Esady and Mate, 2025).

At longer horizons, the Bank's forecasts have continued to be produced using COMPASS (see section 3.2.1) as the central organising framework, supplemented by a broader suite of structural, semi-structural and statistical models. This wider suite helps to capture certain features of the economy that are not particularly well developed within COMPASS, for example the labour market. These suite models also help us to interpret economic dynamics and policy transmission across a range of variables, and provide a range of cross-checks that can help inform staff and MPC judgements more broadly (see Burgess et al., 2013, section 5).

Note, however, that the set of models used to produce the forecast has in the past changed – for example as the Bank moved from BEQM to COMPASS in 2011, or more recently as we have made substantial improvements to COMPASS – and may change again in future. The Bank is continuing to improve its processes and modelling toolkit in response to the Bernanke review (Bernanke, 2024). One implication that should be borne in mind when interpreting results of historical forecast evaluations is that past performance may therefore not always provide a good guide to the future.

3.1.3 Detail on conditioning assumptions

The MPR forecasts are conditional on the following assumptions:

Policy assumptions

- The path for Bank Rate follows the forward Overnight Indexed Swap (OIS) rate curve

- The stock of assets purchased under the MPC's Quantitative Easing (QE) programme evolves in line with market expectations (reflected in asset prices)
- The paths for foreign policy rates follow forward market rates abroad
- UK fiscal policy evolves in line with announced government plans, reflected in OBR projections for government expenditure and taxation

Other financial asset prices

- Wholesale oil and gas prices evolve in line with forward market prices, and non-wholesale energy costs evolve in line with Ofgem projections
- The sterling effective exchange rate index follows a path implied by a 50/50 random walk/uncovered interest parity assumption

Real economy projections

- The UK's working age population is assumed to grow in line with the ONS' population projections
- The UK forecast takes the international outlook (produced by Bank staff) as given (under small open economy assumptions).

The way in which conditioning assumptions are formulated and used can also change over time. For example, the Bank's headline projections changed from being conditional on a constant path for Bank Rate prior to August 2004, to a path based on forward market rates thereafter. Until August 2019, forward market prices were used to pin down oil and gas prices over the forecast. This was replaced with a Random Walk assumption from August 2019 until November 2022, when we returned to using forward market prices as conditioning assumptions for these variables. Moreover, the assumption of a constant stock of purchased assets under the MPC's Quantitative Easing programme, in place since the inception of the programme in March 2009, was relaxed in August 2021.

Further detail on how MPR forecasts are produced, and the role of conditioning assumptions within this, is provided in Section 2 of the Forecast Evaluation Report (Bank of England, 2026).

3.1.4 The role of judgement

Judgement plays a central role in shaping the MPR forecast and enters the process at multiple stages. Early in the process, staff can directly apply or propose judgements. Short-term forecasts, for example, almost always represent staff's best guesses taking into account all available information, rather than mechanical model outputs. Further adjustments, especially concerning longer-term profiles, can also be requested by the MPC as a whole. This process ensures that the forecast is not merely mechanical, but incorporates a significant amount of deliberation and expert insight. Judgements can be particularly important when forecasters possess extra information about real-world events not yet captured in the data, or to correct for known model limitations. For more details, see Burgess et al. (2013).

3.2 Benchmark models

Comparing the performance of the Bank's forecasts to a varied set of models can help gauge the extent to which the Bank's overall forecast process adds value over more mechanical approaches, or whether it could perhaps be improved by leaning more heavily on some particular model, or models. By making a relative comparison, this exercise also helps to control for the impact of changes in economic volatility on forecast errors, since all approaches are subject to this in the same way.

In this section we describe four different model benchmarks, used throughout the paper: a hands-free version of COMPASS, a Bayesian VAR model, and two standard statistical benchmarks in the form of an AR(p) and a random walk model. Moreover, note that the COMPASS and BVAR benchmarks can be run both unconditionally, using the models' endogenous projections of conditioning paths, and conditionally with real-time conditioning assumptions. The ability to run these models conditionally also enables us to use them in counterfactual exercises exploring the role of surprises to conditioning paths, which we discuss further in section 5.

3.2.1 COMPASS

Albuquerque et al. (2025) present the latest iteration of COMPASS – a medium-sized, open economy NK DSGE model – featuring multiple real and nominal rigidities. Such a model provides a coherent framework for understanding macroeconomic relationships and policy effects. The model has a Two-Agent New Keynesian (TANK) structure, featuring fully optimising and rule-of-thumb households. This heterogeneity across agents allows the model to better capture observed economic dynamics in the UK. The model can produce forecasts for a core set of macroeconomic variables that are central to MPC decision-making, including GDP growth and CPI inflation. Compared to previous versions of COMPASS, this latest version includes an enhanced treatment of energy to better capture its role as a key driver of recent inflation dynamics. Conditional forecasts from COMPASS are produced following Burgess et al. (2013, Appendix C).

3.2.2 Bayesian Vector Autoregression (BVAR) model

We use a Bayesian Vector Autoregression (BVAR) model estimated on the same 20 key macroeconomic variables from COMPASS, providing a purely statistical alternative to COMPASS's structural approach. The model specification and estimation follows the work of Giannone et al. (2015) closely, where its high dimensionality is handled with an optimal degree of shrinkage. Covid-19 is accounted for using the approach of Cascaldi-Garcia (2022). Conditional forecasts are produced by constraining future shocks to match the conditional paths following Waggoner and Zha (1999).

3.2.3 Statistical benchmarks

The two models discussed above aim to capture economic relationships and dynamics among a large set of variables. It is also useful to compare the MPR forecast against simpler models that abstract altogether from those wider economic relationships. For

this, we use two common statistical benchmarks: a univariate autoregressive (AR(p)) model and a random walk (RW) model. These models are parsimonious, yet often effective forecasting tools, providing a baseline against which other forecasts can be evaluated.

The lag order p of the AR(p) model is selected using the Bayesian Information Criterion (BIC) from a maximum of two lags. These AR models are fitted using Maximum Likelihood Estimation (MLE) with location-scale t-distributed errors, which allows for robust handling of outliers and heavy-tailed distributions. More details on how these are estimated can be found in appendix A.2. The random walk model simply assumes that the best prediction is the last observed value, so no estimation is required.

3.3 Variables, estimation, and real-time settings

The forecast evaluation techniques described in this paper are illustrated through selected applications to four variables relevant to monetary policymakers: year-on-year GDP growth, in-house seasonally adjusted CPI inflation (henceforth referred to as CPI inflation), wage growth, and the level of the unemployment rate. These variables are also the focus of the 2026 Forecast Evaluation Report (Bank of England, 2026), and included in the accompanying Python package. We generally evaluate forecasts from 2015. We analyse forecasts up to twelve quarters ahead, starting with the current quarter.

All benchmark models are estimated using quarterly data, the same frequency of the MPR forecasts. Statistical benchmarks and the BVAR are estimated from 1997Q3. COMPASS is estimated from 1987Q4, with potential structural breaks over that longer period handled through the use of time-varying trends.

As economic data can be revised over time to reflect new information and methodological improvements, using real-time data is crucial to ensure meaningful forecast comparisons. Each model is therefore recursively estimated and used to project the key variables out-of-sample using the same information set that was available in real-time at each point of the 2015-25 evaluation window. This involves using the appropriate vintages of back data and conditioning assumptions for each time period.

4 Evaluating historical forecast errors statistically

This section discusses the array of statistical evaluation metrics and techniques that we use to assess the historical performance of the Bank of England's forecasts along three traditional dimensions: accuracy, bias and efficiency. We illustrate each through selected applications to the Bank's MPR forecasts of four key macroeconomic variables: GDP growth, the unemployment rate, wage growth, and CPI inflation.

We begin with some key definitions and notation in section 4.1. We then explore the following properties in sections 4.2-4.4:

- **Accuracy:** How close to economic outcomes have MPR forecasts typically been, and have they been more precise than benchmarks on average?

- **Unbiasedness:** Have MPR forecasts tended to over- or under-predict economic outcomes?
- **Efficiency:** Have MPR forecasts made optimal use of information available?

Worked examples shown through this section generally cover the period from 2015Q1 to 2025Q4, unless stated otherwise.

4.1 Notation

We work with quarterly data which is the frequency of the Bank's forecast. Each vintage is named according to the month and year of its release. For example, a forecast released in November 2011 is referred to as N11, while a forecast released in February 2012 is referred to as F12. The naming convention follows the pattern where F=February, M=May, and A=August and N=November. We continue to use this notation throughout the paper, and put vintage names in parentheses when referring to the quarter they are published in, eg (N11) refers to 2011Q4.

We distinguish between two types of observations: forecasts and outturns. Outturns refer to published data, which can be subject to revision and thus change from one vintage to the next. Each vintage contains both forms of observations: forecasts for the current quarter – the 'nowcast' – and the next 12 quarters, and outturns for the previous quarters.

For each variable y, we denote an observation of quarter t in vintage v as $y_{t|v}$. For example, $y_{2012Q2|(N11)}$ is the value for variable y in 2012Q2 taken from the November 2011 vintage.

There is a clear separation between forecasts and outturns: $y_{t|v}$ is a forecast if $v \leq t$ and an outturn if $v > t$. In order to highlight forecasts, we denote such observations with a hat, ie we refer to forecasts as $\hat{y}_{t|v}$ when $v \leq t$. Also note that a nowcast is a forecast for the same quarter as the vintage is published, that is $\hat{y}_{t|t}$.

Forecast errors are defined as the difference between the outturn and the forecast. Because outturns are subject to revision across vintages, the forecast error can vary depending on which vintage of data is used for the comparison. We define the h-quarter ahead forecast error for variable y for a value in quarter t, based on published data k quarters after the first release as:

$$\varepsilon(y;k)_{t|t-h} \coloneqq y_{t|t+1+k} - \hat{y}_{t|t-h}.$$

The parameter k determines which vintage of the outturn is used to compute forecast errors. Setting $k = 0$ uses the earliest available data (available one quarter after the reference period), $k = 1$ uses the first revision, and so on. Where series are subject to revision, choosing k involves a trade-off: larger k yields more stable, revised outturns (improving data quality) but increases the chance that series definitions or compilation methods have changed, reducing comparability with the original forecasts.

We set $k = 12$ throughout the paper, meaning that we compare forecasts against the outturn published 13 quarters after the reference period. This ensures that GDP data

Table 1: Forecasts and outturns for CPI inflation

	N21	F22	M22	A22	N22	F23	M23	A23	N23	F24	M24	A24	N24	F25	M25	A25	N25
2021:Q4	4.33	4.91	4.91	4.90	4.90	4.88	4.89	4.90	4.91	4.91	4.90	4.91	4.91	4.91	4.91	4.91	4.91
2022:Q1	4.56	5.73	6.22	6.22	6.24	6.27	6.25	6.26	6.27	6.28	6.26	6.27	6.28	6.28	6.30	6.30	6.30
2022:Q2	4.80	7.03	9.12	9.16	9.16	9.14	9.14	9.12	9.11	9.09	9.14	9.12	9.12	9.11	9.10	9.10	9.10
2022:Q3	4.42	6.70	9.46	9.93	10.02	10.02	10.04	10.04	10.03	10.03	10.01	10.01	10.00	10.01	10.00	10.00	10.00
2022:Q4	3.40	5.81	10.19	13.10	10.86	10.74	10.74	10.74	10.75	10.75	10.75	10.76	10.76	10.77	10.77	10.78	10.78
2023:Q1	3.26	5.21	9.31	12.54	10.13	9.76	10.19	10.20	10.22	10.22	10.21	10.22	10.23	10.23	10.25	10.26	10.26
2023:Q2	2.58	3.49	6.65	10.78	9.54	8.46	8.20	8.40	8.39	8.37	8.40	8.39	8.39	8.37	8.35	8.34	8.34
2023:Q3	2.40	3.25	5.88	9.52	7.87	6.20	6.98	6.95	6.72	6.72	6.70	6.70	6.70	6.70	6.69	6.70	6.70
2023:Q4	2.23	2.45	3.56	5.46	5.20	3.92	5.12	4.93	4.63	4.17	4.18	4.19	4.19	4.19	4.20	4.21	4.21
2024:Q1	2.11	2.15	2.91	4.33	3.96	3.01	4.41	4.30	4.39	3.61	3.56	3.56	3.58	3.59	3.59	3.59	3.59
2024:Q2	1.99	1.89	2.14	2.64	1.09	0.96	3.38	3.27	3.63	1.95	1.98	2.08	2.07	2.05	2.06	2.04	2.04
2024:Q3	1.96	1.77	1.85	2.00	1.16	1.67	2.91	2.78	3.30	2.19	2.21	2.27	2.03	2.04	2.03	2.03	2.03
2024:Q4	1.95	1.66	1.51	1.40	1.43	1.42	2.28	2.48	3.13	2.65	2.58	2.73	2.36	2.48	2.48	2.49	2.48
...		...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

This table shows forecasts and outturns for CPI inflation for the period 2021Q4 to 2025Q4. Each column represents a vintage from N21 to N25, while each row represents either an outturn or a forecast made at that vintage for a given quarter. The numbers coloured in blue correspond to the nowcast, while the four-quarter ahead forecast is coloured in orange, the eight-quarter ahead forecast is coloured in cyan and the 12-quarter ahead forecast is coloured in magenta. We colour the relevant outturns in red. The outturns are defined as data from $k = 12$ quarters after the first release of data. If the outturns data are not yet available, we take the outturn from the latest available vintage. For this example, we assume that N25 is the latest vintage.

in particular will have been fully 'balanced'[1] at least twice in the ONS's annual Blue Book publication, where sources and methods used for UK National Accounts data are revisited comprehensively. Whenever the $t + k + 1$ vintage has not yet been released, we use outturns from the latest available vintage. To simplify notation, we suppress the vintage subscript in what follows. That is, we define the outturn of variable y in period t as $y_t := y_{t|t+1+12}$ and the forecast error for variable y as:

$$\varepsilon(y)_{t|t-h} := y_t - \hat{y}_{t|t-h}. \tag{1}$$

Example for CPI inflation: Table 1 shows CPI inflation forecasts and outturns for the period 2021Q4 to 2025Q4. Rows denote time periods of observations, that is the period of the outturns or forecasts, while columns denote vintages. The November 2021 (N21) forecast is the first vintage shown. For this forecast, data up to 2021Q3 is available. The forecast for the current calendar quarter, which we generally refer to as the 'nowcast', is shown in blue. Forecasts extend 12 quarters into the future. Outturns from the vintage $k = 12$ quarters after the first release of data are highlighted in red. If the outturns data are not yet available, we take the outturn from the latest available vintage instead.

A popular way to visualise forecast vintages and outturns together are so-called 'hedgehog' plots. For instance, fig. 1 shows the hedgehog plot of CPI inflation forecasts from the MPR. Deviations between the forecasts and the outturns represent forecast errors.

We now turn to describing the evaluation metrics and techniques used in the toolkit.

[1]GDP estimates are computed using three different approaches: production, income and expenditure. Each approach makes use of distinct data and the three resulting estimates are reconciled, or balanced, in the ONS Blue Book (Office for National Statistics, 2019).

4.2 Accuracy

The usual first step in learning from historical forecasts consists in understanding their accuracy. We differentiate between absolute accuracy and relative accuracy. Absolute accuracy measures how close forecasts are to the outturns. Relative accuracy compares the forecasting performance of the MPR to the benchmark models described in section 3. Benchmarks provide a minimum forecasting performance while controlling for changes in realised economic volatility.

4.2.1 Accuracy metrics

We implement two measures of absolute accuracy: the root mean squared error (RMSE) and the mean absolute error (MAE). Let $\{\varepsilon_i\}_{i=1}^N$ denote a sample of N forecast errors defined in eq. (1). The RMSE is defined as:

$$\text{RMSE} := \sqrt{\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i^2}, \tag{2}$$

while the MAE is defined as:

$$\text{MAE} := \frac{1}{N}\sum_{i=1}^{N}|\varepsilon_i|. \tag{3}$$

The MAE is less sensitive to outliers and is more easily interpretable as a simple average of absolute forecast error sizes, whereas the RMSE places more weight on larger errors. It is also possible to abstract further from outliers by computing the equivalent metrics using the median instead of the mean, as in Kanngiesser and Willems (2024). For the remainder of the paper we use the RMSE as the primary accuracy metric, for consistency with the quadratic loss function that forecasters typically seek to minimise.

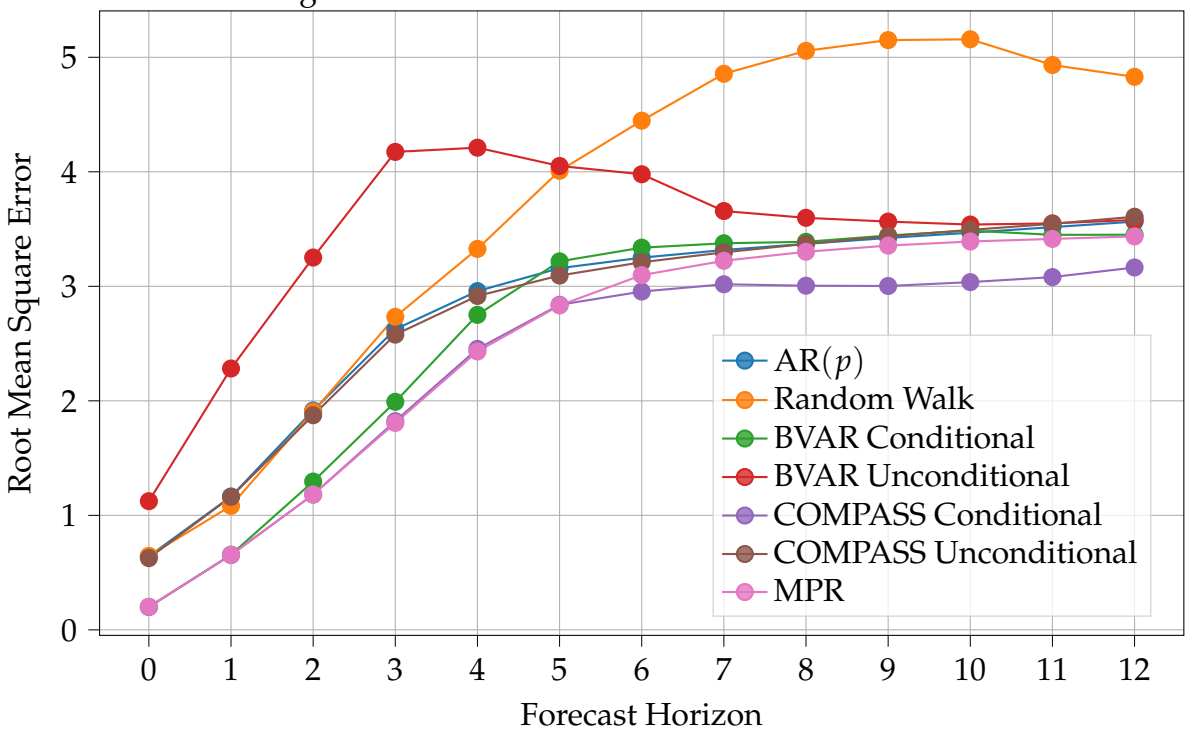
Example for CPI inflation: Figure 2 shows the RMSEs for the MPR forecasts of CPI inflation compared to the set of benchmark models. The MPR forecast performs at least as well as each of the benchmarks at most forecast horizons, although the conditional COMPASS forecasts have a slightly lower RMSE at longer horizons. Generally, we see that the RMSE increases with the forecast horizon, which is expected as outturns become harder to predict the further into the future we go.

4.2.2 Relative accuracy and the Diebold-Mariano test

Comparing a set of forecasts with a measure of accuracy like RMSE or MAE gives a first insight into their relative forecast performance. But analysing the relative accuracy of two given forecasts can also be done more directly by taking the ratio of their RMSEs:

$$\text{RMSE ratio} := \frac{\text{RMSE}_{\text{Forecast A}}}{\text{RMSE}_{\text{Forecast B}}}, \tag{4}$$

Figure 2: RMSEs for forecasts of CPI inflation



This figure shows the RMSE for the h-quarter ahead forecasts for CPI inflation. The sample period is from 2015Q1 to 2025Q4.

A ratio greater than one indicates that the denominator, Forecast B, performs better, and vice versa. Given its intuitive interpretation, this is the measure we report in the paper, where the MPR forecast is used as the denominator.

The RMSE ratio, however, does not indicate whether the differences in accuracy are statistically significant. To investigate statistical significance we use the Diebold-Mariano test of relative predictive accuracy (Diebold and Mariano, 1995). The test evaluates the null hypothesis that the expected difference of forecast loss – typically measured by squared errors – is zero, implying that both models exhibit equal predictive accuracy. Although the difference of squared errors and the ratio of RMSEs are not strictly equivalent measures of relative accuracy, they rank models identically.

We use a Heteroscedasticity and Autocorrelation Consistent (HAC) variance estimator to account for autocorrelation in the loss differential series when evaluating forecasts at horizons greater than one. Following Harvey et al. (2017), we use the original variance estimator of Diebold and Mariano (1995) by default and only use the Bartlett estimator if the original estimator yields a negative variance; the original estimator has better properties for small samples. We also use the Harvey et al. (1997) correction to account for small sample sizes.

The resulting test statistic and p-value indicate the probability of observing such a difference under the null hypothesis of equal forecast performance. A positive test statistic indicates the base model performs better, whereas a negative test statistic indicates that it performs worse.

Table 2: Relative accuracy and Diebold-Mariano test results for CPI inflation

	Forecast horizon					
Model	0	1	2	4	8	12
AR(p)	3.21	1.78	1.62	1.22	1.02	1.04
BVAR Conditional	1.00	1.00	1.10	1.13	1.03	1.00
BVAR Unconditional	5.61	3.49	2.75	1.73	1.09	1.04
COMPASS Conditional	1.00	1.00	1.00	1.01	0.91	0.92
COMPASS Unconditional	3.13	1.78	1.59	1.20	1.02	1.05
Random Walk	3.22	1.66	1.61	1.37	1.53	1.41
Observations	43	42	41	39	35	31

This table shows the ratio of RMSEs from forecasts of CPI inflation from various models to the MPR's forecasts at different forecast horizons. Coloured cells indicate a statistically significant difference (p-value < 0.05) in predictive accuracy compared to the MPR, based on the Diebold-Mariano test. A value greater than one implies that the MPR's forecasts are more accurate than the benchmark model (shaded green), while a value less than one indicates MPR's forecasts are less accurate (shaded red). The number of observations varies by forecast horizon due to the availability of data.

Note that the Diebold-Mariano test assumes that the loss differential series has constant mean, variance and autocovariance. The results can be misleading when these assumptions are violated, for example because of instabilities like structural breaks or temporary volatility. Hence large differences in average forecasting performance may not appear significant for this reason. Section 5 discusses an attempt to mitigate the effect of instabilities – which affect tests in general – by using fluctuation tests.

Example for CPI inflation: Table 2 shows relative accuracy and Diebold-Mariano test results for CPI inflation. The table shows that the MPR exhibits consistent accuracy gains over the models at most forecast horizons, particularly at short to medium horizons. At longer horizons, however, the conditional COMPASS forecasts perform slightly better, suggesting that the model-based forecasts can be competitive with the MPR when projecting further ahead. Overall, these results indicate that the MPR forecast process described in the previous section – which combines multiple models with expert judgement – adds significant value.

4.3 Unbiasedness

A forecast is biased if it systematically over- or under-predicts the outturn. Equivalently, if bias is present, forecast accuracy could have been improved by adding a constant adjustment. We test for bias by regressing forecast errors on a constant and assessing whether the implied average error differs significantly from zero:

$$y_{t|t-h} = \beta + u_t, \tag{5}$$

where β is a scalar coefficient and u_t is a zero-mean error term. The OLS estimate of $\hat{\beta}$ is the sample mean of the forecast errors. A $\beta > 0$ implies that forecasts, on average, underestimate the outturn, while $\beta < 0$ implies that they overestimate it.

Table 3: Comparison of bias results for GDP growth

	Forecast horizon					
Model	0	1	2	4	8	12
AR(p)	0.12	-0.12	-0.37	-0.57	-0.68	-0.77
BVAR Conditional	0.31	-0.07	-0.50	-1.16	-0.51	-0.43
BVAR Unconditional	0.21	0.23	0.34	0.86	-1.51	-1.26
COMPASS Conditional	0.31	0.15	0.06	-0.35	-0.59	-0.52
COMPASS Unconditional	0.28	0.24	0.32	0.32	0.10	0.03
MPR	0.31	0.24	0.27	0.08	-0.09	-0.28
Random Walk	0.24	0.11	-0.03	-0.58	-0.98	-1.39
Observations	35	33	32	30	26	22

The table reports the estimated bias coefficients (in percentage points) for the forecasts of GDP growth from various models at various forecast horizons. The null hypothesis is that the forecasts are unbiased. Cells with statistically significant bias (p-value < 0.05) are highlighted in red. The number of observations varies by forecast horizon due to the availability of data. Forecast errors from the pandemic period (2020Q1 to 2022Q1) have been excluded from the analysis.

The null hypothesis is that the forecasts are unbiased, $\beta = 0$. We use a t-test to test this hypothesis. The test statistic is given by:

$$t = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})}, \tag{6}$$

where $\hat{\beta}$ is the estimated coefficient and $\text{SE}(\hat{\beta})$ is the standard error of the estimate.

We fit the regression model using Ordinary Least Squares (OLS) with HAC standard errors. We assume a maximum lag of h quarters for the HAC standard errors.

Example for GDP growth: Table 3 shows bias test results for forecast of GDP growth for the MPR as well as for the benchmark models. The results indicate that the MPR's forecasts are unbiased at most forecast horizons, with the exception of nowcasts, which at face value appear to underestimate outturns. As discussed in the Forecast Evaluation Report (Bank of England, 2026), this particular result is driven by exceptionally large upward revisions to GDP growth data in the post-pandemic period.

4.4 Efficiency

A forecast is efficient if it cannot be improved upon given the information available at the time of the forecast. Economists generally work with two types of efficiency, defined by Nordhaus (1987): strong efficiency and weak efficiency.

A strongly efficient forecast cannot be improved upon when using all information available at the time of the forecast. The difficulty in defining and obtaining this information set is a barrier to testing conclusively for strong efficiency in practice.

On the other hand, a forecast is weakly efficient if it cannot be improved upon when using its own history only, ie past forecasts. It is safe to assume that a forecaster has access to this information.

4.4.1 Testing weak efficiency

Weakly efficient forecasts should satisfy the following three conditions (Independent Evaluation Office, 2015):

- **No systematic bias**: Forecasts should not consistently over- or under-predict in ways that could be corrected by simple adjustments to the forecasts.
- **No predictable pattern**: Past forecast revisions should not predict future revisions. When they do, it indicates information smoothing, which is the gradual incorporation of news over multiple forecast rounds rather than immediate adjustment. Efficient forecasters should respond decisively to new data rather than spreading adjustments across multiple vintages.
- **Optimal information weighting**: Forecasts should incorporate new information optimally, neither overreacting nor underreacting. Overreaction occurs when forecasters put too much weight on recent data, leading to volatile forecast revisions that overshoot the truth. Underreaction happens when forecasters are too anchored to previous views, updating insufficiently in response to genuinely informative signals.

In practice, predictability and suboptimal weighting co-occur because they stem from incomplete information processing (Coibion and Gorodnichenko, 2015). When forecasters smooth information, they simultaneously create predictable revision patterns and reveal suboptimal weighting of new data.

1. Testing bias and optimal scaling conjointly: the Mincer-Zarnowitz Regression

We first test whether forecasts could have been made more accurate by adding a constant term or scaling them by a constant factor. We do this using the following regression model for optimal scaling (Mincer and Zarnowitz, 1969):

$$y_{t+h} = \beta_0 + \beta_1 \hat{y}_{t+h|t} + u_{t+h}, \tag{7}$$

where we regress the outturn of variable y on a constant and the h-quarter ahead forecast. u is a zero-mean error term.

We fit the regression model using OLS with HAC standard errors. We assume a maximum lag of h quarters for the HAC standard errors.

The null hypothesis is that $\beta_0 = 0$ and $\beta_1 = 1$ and we use an F-test to test this joint hypothesis.

Example for GDP growth: Table 4 compares optimal scaling test results for MPR and benchmark forecasts of GDP growth. Results suggest that the MPR's GDP forecast are not weakly efficient at the majority of forecast horizons. The same applies to most

Table 4: Comparison of optimal scaling results for GDP growth

Model	Forecast horizon 0	1	2	4	8	12
AR(p)	0.18	0.15	0.00	0.00	0.01	0.00
BVAR Conditional	0.05	0.01	0.00	0.00	0.03	0.00
BVAR Unconditional	0.12	0.00	0.00	0.00	0.00	0.00
COMPASS Conditional	0.05	0.12	0.13	0.00	0.00	0.00
COMPASS Unconditional	0.03	0.02	0.00	0.27	0.89	0.00
MPR	0.05	0.03	0.00	0.00	0.09	0.04
Random Walk	0.16	0.00	0.00	0.00	0.00	0.00
Observations	35	33	32	30	26	22

The table reports the p-values from the optimal scaling test for the forecasts of GDP growth from various models at various forecast horizons. The null hypothesis is that the forecasts are optimal. P-values less than 0.05 are highlighted in red. The number of observations varies by forecast horizon due to the availability of data. Forecast errors from the pandemic period (2020Q1 to 2022Q1) have been excluded from the analysis.

benchmark models, however, suggesting that optimal scaling is a common issue across models, potentially due to the prevalence of frequent revisions in GDP data.

2. Testing predictability of forecast revisions

If forecasters process information efficiently, each forecast revision should reflect only new information that was unavailable in previous periods. When past revisions predict future revisions, it reveals information smoothing, which is the gradual incorporation of news across multiple forecast rounds rather than immediate adjustment.

We test whether revisions are predictable following the approach of Nordhaus (1987), which was also applied to the Bank's forecast in Independent Evaluation Office (2015). We define the forecast revision for variable y at time t from vintage $v \leq t$ as:

$$R(y)_{t|v} := \hat{y}_{t|v} - \hat{y}_{t|v-1}, \tag{8}$$

where $\hat{y}_{t|v}$ is the forecasted value of variable y of quarter t taken from vintage v, and $\hat{y}_{t|v-1}$ is the forecasted value of the same quarter t, but taken from the previous vintage $v-1$. For example, $R(y)_{t|t}$ would denote the final forecast revision of the observation in quarter t. Since our forecasts extend from the nowcast up to 12 quarters ahead, there are in total 12 forecast revisions for any forecasted variable y at time t.

To test whether the final forecast revision is predictable, we regress the final forecast revision on earlier revisions:

$$R(y)_{t|t} = \alpha + \sum_{i=1}^{N} \beta_i R(y)_{t|t-i} + u_t, \tag{9}$$

where we consider N earlier revisions, and u_t is a zero-mean error term. We choose to include $N = 5$ past forecast revisions as predictors, focusing on the most recent

Table 5: Forecast revision predictability results for the unemployment rate

Model	F-statistic	P-value	Observations
AR(p)	6.97	0.00	38
Random Walk	8.06	0.00	38
MPR	20.29	0.00	38

This table presents results from F-tests of the joint significance of lagged revision coefficients in forecast revision regressions for the unemployment rate. The null hypothesis is that all lagged revision coefficients are zero (revisions are unpredictable). Rejecting the null suggests forecast revisions are predictable, indicating inefficient information processing. P-values less than 0.05 are highlighted in red. HAC standard errors are used to account for serial correlation and heteroskedasticity. The same date range is used for all models to ensure fair comparison.

revisions (earlier revisions relate to older news, which are less likely to be relevant for predicting the final revision). The null hypothesis is that all $\beta_i = 0$, indicating that the final forecast revision is not predictable from past forecast revisions. We use an F-test to test this joint hypothesis.

Example for unemployment: Table 5 shows the forecast revision predictability test results for MPR and available model-based forecasts of the unemployment rate, which in this case are only the two simple statistical benchmarks. The results suggest that the MPR's forecast revisions for the unemployment rate are predictable from past revisions at the 5% significance level.

3. Testing correlation between forecast revisions and forecast errors

Finally, we test whether forecast revisions are correlated with forecast errors, which would be an indication of sub-optimal information weighting (see Coibion and Gorodnichenko (2015) for a recent discussion on this approach). We use the following regression model:

$$\varepsilon(y)_{t+h|t} = \alpha + \beta R(y)_{t+h|t} + u_{t+h}, \tag{10}$$

where $\varepsilon(y)_{t+h|t}$ is the h-quarter ahead forecast error for variable y, the term $R(y)_{t+h|t}$ represents the forecast revision of the forecast with an eventual h-quarter-ahead forecast horizon, and u_t is a zero-mean error term. The null hypothesis is that $\beta = 0$, indicating that forecast revisions are uncorrelated with forecast errors.

A $\beta > 0$ suggests that forecasters are systematically underreacting to new information by under-revising their forecasts. Conversely, $\beta < 0$ indicates that forecasters are systematically overreacting to new information by over-revising their forecasts. We use a t-test to test this hypothesis. The regression model is fitted using OLS with HAC standard errors to account for potential heteroscedasticity and autocorrelation in the forecast errors. We assume a maximum lag of h quarters for the HAC standard errors.

Example for CPI inflation: Table 6 shows the correlation between forecast revisions and forecast errors for the MPR's forecasts of CPI inflation. The results indicate that the

Table 6: Correlation between forecast revisions and forecast errors for CPI inflation

Horizon	$\hat{\beta}$	SE($\hat{\beta}$)	P-value	Observations
0	0.06	0.05	0.26	42
1	0.02	0.23	0.93	41
2	0.20	0.25	0.42	40
3	0.33	0.35	0.35	39
4	0.21	0.43	0.62	38
5	0.52	0.54	0.34	37
6	-0.11	0.42	0.80	36
7	-0.31	0.80	0.70	35
8	-1.51	2.49	0.54	34
9	-2.67	3.57	0.45	33
10	-4.46	2.65	0.09	32
11	-2.33	2.14	0.27	31

This table shows the correlation between forecast revisions and forecast errors test results for the MPR's forecasts of CPI inflation. The null hypothesis is that forecast revisions are uncorrelated with forecast errors. P-values less than 0.05 are highlighted in red. HAC standard errors are used to account for serial correlation and heteroskedasticity. The number of observations varies by forecast horizon due to the availability of data.

forecast revisions and forecast errors for CPI inflation are uncorrelated at all forecast horizons.

4.4.2 Strong efficiency

Strong efficiency tests whether forecast errors can be predicted using the full information set available at the time the forecast was made. In practice, however, it is not feasible to include the complete information set in our tests. We therefore restrict the information set to forecasted values of other variables produced at the same time, conducting a series of bivariate tests.

For these bivariate tests, we use Blanchard-Leigh regressions, following the approach in Kanngiesser and Willems (2024). This method was first developed in Blanchard and Leigh (2013) to analyse whether the IMF's forecasts of the relationship between planned fiscal consolidation and economic growth were accurate. In general, this type of regression investigates whether forecasts systematically over- or under-estimate the relationship between one variable and another. For instance, in a monetary policy setting, forecasters often have to rely on certain assumptions about how GDP growth affects CPI inflation (demand-pull effects), how wage growth feeds into consumer prices (cost-push channels), or how unemployment influences wage pressures (Phillips curve dynamics). These relationships are also embedded in the Bank's forecasting models, whether explicitly through structural equations or implicitly through reduced-form correlations learned from historical data.

To test the calibration of selected forecast relationships, we adopt the extension of Kanngiesser and Willems (2024) that corrects for potential systematic errors in the

forecasted variable:

$$\varepsilon(y)_{t+h|t} = \alpha + \beta \hat{x}_{t+j|t} + u_{t+h}, \tag{11}$$

$$x_{t+j} = \gamma + \delta \hat{x}_{t+j|t} + e_{t+j}, \tag{12}$$

where $\hat{x}_{t+j|t}$ is the j-quarter ahead forecasted value for variable x and x_{t+j} is the outturn of the variable. u_t and e_t are zero-mean, potentially correlated error terms. Note that the forecast of interest y and the explanatory variable x are from the same vintage t, but we allow for different forecast horizons, j and h. The regression coefficient β measures the relationship between the forecast errors of variable y and the forecasted values of variable x.

Equation (12) is a Mincer-Zarnowitz regression, which corrects for systematic errors in the forecast of variable x. The regression coefficient δ measures the relationship between the outturn and the forecasted values of variable x. If $\delta \neq 1$, then the forecasts of variable x contain some systematic error.

We are interested in estimating the Wald ratio, $\omega(\beta, \delta)$ which is defined as:

$$\omega(\beta, \delta) = \frac{\beta}{\delta}. \tag{13}$$

The Wald ratio tells us whether the Bank's forecasts over or under-estimate the 'pass-through' of variable x to variable y at forecast horizon h. More precisely, it captures whether forecast errors in y are systematically related to forecasted values of x in a way that could have been anticipated and corrected.

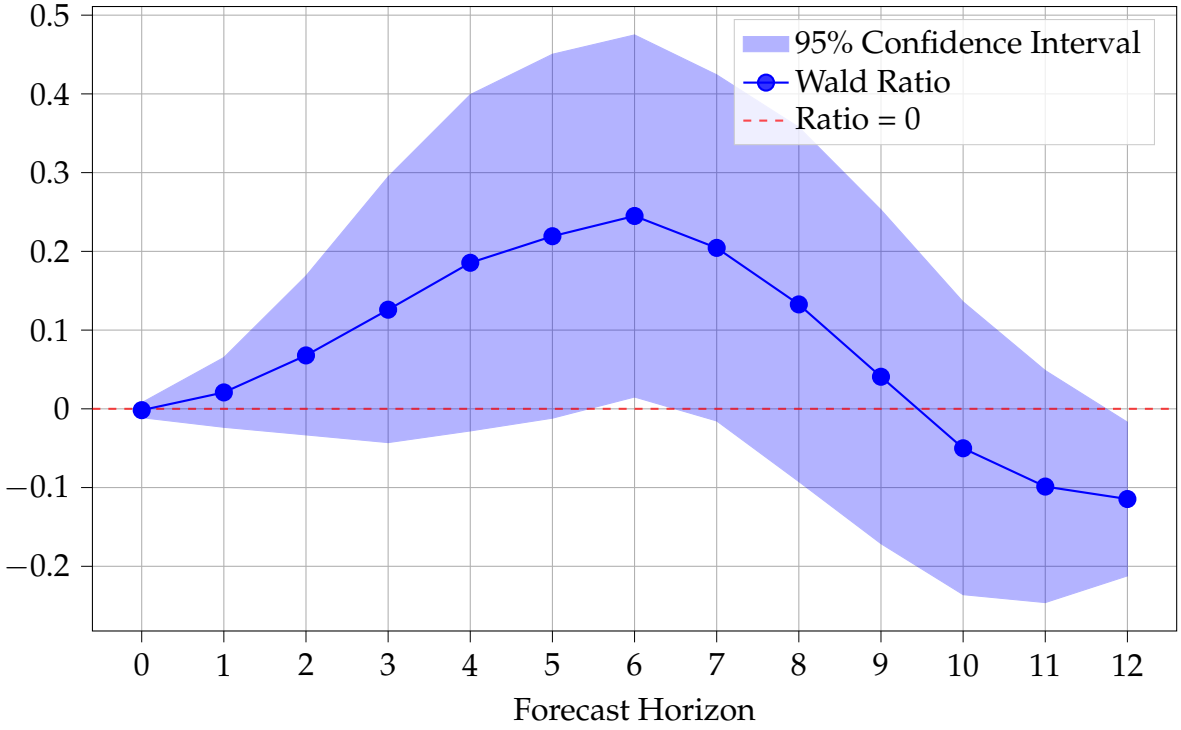
A positive estimate indicates that, on average, higher forecasted values of variable x are followed by higher-than-forecast outturns of y. This suggests that the model-implied 'pass-through' of variable x to variable y might be too low as the forecast failed to fully account for how increases in x would be reflected in higher y. For example, if testing CPI inflation against GDP growth, a positive Wald ratio would suggest that forecasts underestimated how stronger growth would push up on inflation.

Conversely, a negative estimate would indicate a potential overestimation of the pass-through: the forecast anticipated a stronger relationship than actually materialised. Using the same example, this would mean the forecasts overestimated how GDP growth would boost CPI inflation, perhaps by overstating demand pressures.

A zero Wald ratio indicates efficient forecasts that capture the relationship correctly on average, neither systematically under- nor over-estimating the pass-through of variable x to variable y.

We take $j = 2$ in both equations, which means that we use the 2-quarter ahead forecast of variable x as the proxy for the outturn of variable x. This choice balances the need to rely on information available to the forecaster at the time (realised outturns are not known when the forecast is made) with the need for the forecast of x to be informative (at longer horizons, forecasts become noisier). The two equations, eq. (11) and eq. (12), are estimated jointly using a system of seemingly unrelated regressions (SUR) with OLS, allowing us to account for potential correlation between the error terms u and

Figure 3: Blanchard-Leigh: CPI inflation forecast errors against GDP growth forecasts



This figure shows the results from the MPR's Blanchard-Leigh regressions of CPI inflation forecast errors on the 2-quarter ahead forecasts of GDP growth. The blue line represents the Wald ratio at each forecast horizon, with the shaded area indicating the 95% confidence interval. A ratio significantly different from zero indicates forecast inefficiency.

e. Standard errors are computed using HAC estimators with a maximum lag of h quarters. While our approach incorporates the SUR framework and HAC standard errors, accounting for serial and cross-equation correlation within the standard errors, Kanngiesser and Willems (2024) adopt a robust regression method to mitigate the influence of outliers, focusing instead on a setup with two single-equation regressions.

We test the null hypothesis that the forecasts are efficient, which is equivalent to testing whether $\omega(\beta, \delta) = 0$. We use a two-sided Z-test to check this hypothesis. The test statistic is given by:

$$z = \frac{\widehat{\omega(\beta, \delta)}}{\text{SE}(\widehat{\omega(\beta, \delta)})}, \tag{14}$$

where $\widehat{\omega(\beta, \delta)}$ is the estimated Wald ratio and $\text{SE}(\widehat{\omega(\beta, \delta)})$ is the estimated standard error of the Wald ratio. Standard errors are computed using the delta method as outlined in appendix A.3.

Figure 3 shows the estimated Wald ratios from the Blanchard-Leigh regressions of CPI inflation against GDP growth. The figure shows that for forecasts up to nine-quarters ahead, higher forecasts for GDP growth are associated with CPI inflation outturns that are higher than forecasted. For forecasts beyond that, the opposite is true.

5 Learning from recent forecast errors

In the previous section we established a toolkit of statistical evaluation techniques to assess historical forecast performance. Complementary to that historical analysis, the focus of this section is on methods that help shed light on forecast errors and their drivers at higher frequency and in real-time. This poses a distinct set of challenges. With recent errors, we do not have a long historical sample to draw on, making it difficult to apply statistical tests with sufficient power. To draw timely conclusions and identify emerging forecast performance issues, we therefore need more targeted approaches. These approaches require a more flexible toolkit, adapting as needed to the specific questions, variables and forecast horizons of interest. Here we do not aim to be exhaustive, but rather provide examples of these broader approaches focusing on certain aspects of the period since mid-2021, which saw sharp rises in energy prices followed shortly after by increases in Bank Rate.

In light of these challenges, section 5.1 outlines exercises for detecting potential issues in real time. First, we examine whether errors are unusually large relative to history and therefore warrant further investigation. Second, we evaluate statistical evidence of any shifts in the long-run historical characteristics of the Bank's forecast errors. Third, we examine the extent to which our assessment of forecast errors might change as data get revised. Each exercise looks at a different macroeconomic variable, owing to the idiosyncratic factors affecting these data and consequently the forecast errors.

Section 5.2 then demonstrates how we can use economic models to help identify the sources of economic surprises and inform real-time views on the outlook. These models help us consider the extent to which recent forecast errors may have been caused by identifiable factors such as developments in forecast conditioning assumptions, versus other factors which could include misspecification of certain structural aspects of the economy or the propagation of past shocks. We present examples of counterfactual exercises and historical decompositions that help us to interrogate these drivers.

5.1 Detecting weaknesses in real time

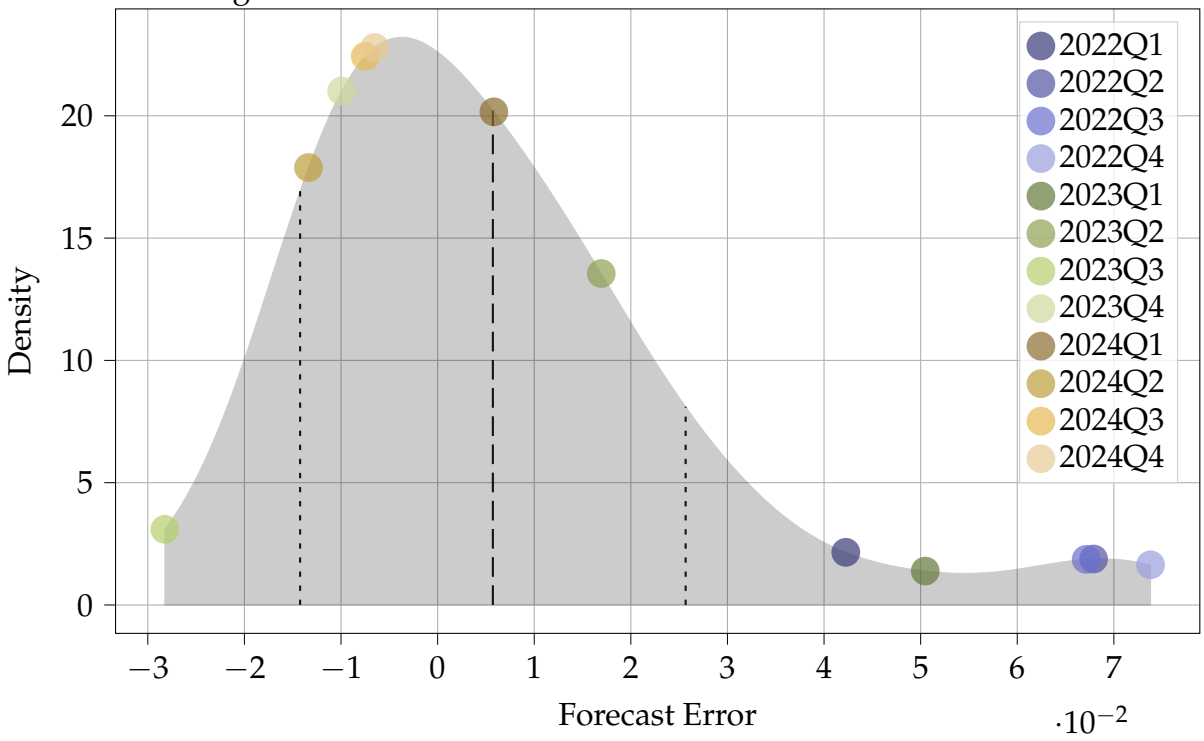
To identify potential weaknesses in our forecasting process as they emerge, we employ a suite of diagnostic tools that monitor forecast performance in real time. These can help us distinguish between routine data variance, atypical patterns that may simply reflect unforecastable exogenous shocks, and more fundamental issues that may require a reassessment of the Bank's models or judgements.

Identifying deviations from historical performance

First, establishing whether a forecast error is typical or atypical relative to historical norms provides an essential starting point for deeper investigation.

There are many ways of considering and visualising this. One is to compare the size of particular errors of interest to longer-term accuracy measures such as the RMSEs discussed in section 4.2, which capture the size of a 'typical' error over the relevant comparison period. The forecast error distribution is another example of a simple device

Figure 4: Distribution of forecast errors for CPI inflation



This figure shows MPR's recent observations alongside the density of 4-quarter ahead forecast errors of CPI inflation. The long dashed line represents the mean forecast error, while the short dashed lines indicate one standard deviation from the mean. The distribution of forecast errors is based on forecasts from 2006Q3 to 2025Q3.

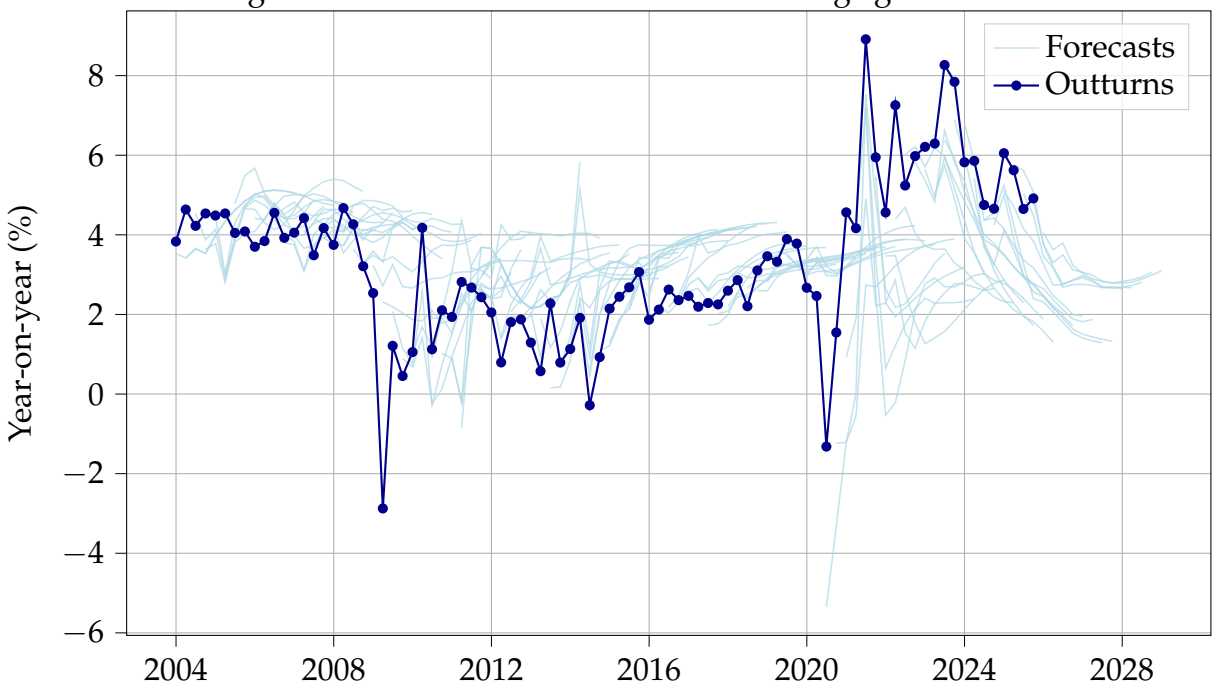
that can be used to place errors in their historical context. Errors falling in the tails of the distribution indicate unusually large deviations from average performance that may warrant further examination. Figure 4 shows the distribution of one-year-ahead forecast errors for CPI inflation. Errors in 2022 (ie the purple circles) are concentrated in the right tail of the distribution, highlighting the unprecedented scale of the extent to which CPI inflation surprised to the upside in that year, relative to previous forecasts. Another error from 2023Q1 also falls in the right tail. However, by 2024 forecast errors were back within the historical norm.

Working with structural breaks and changing environment

Second, we investigate how we might identify changes in forecast performance based on the latest errors.

The statistical tests in section 4 are a useful first step for distinguishing meaningful patterns in the data from chance, but have limitations, including an implicit reliance on a stable environment where the properties of the data generating processes do not change over time. For example, the Diebold-Mariano test assumes that the difference in forecast loss between two models has constant mean and variance. In practice, this assumption may fail due to structural breaks in the data or evolving economic relationships between variables.

Figure 5: MPR forecasts and outturns of wage growth



The light blue lines show forecasts from different vintages. The dark blue line shows the outturns $k = 12$ quarters after the first data release.

Therefore, relying on full sample statistical tests when there is emerging evidence of structural breaks and a departure from historical patterns can mask important changes in forecast performance. In such cases it is more informative to use rolling-window analysis supplemented with fluctuation tests, which can better handle small samples.

A rolling window approach can mitigate the effect of instability by focusing on shorter periods but also introduces drawbacks: smaller samples reduce power, and repeated testing increases the risk of false rejections of null hypotheses. To account for this, we follow Giacomini and Rossi (2010) who propose a formal testing approach for rolling windows – the 'fluctuation test' – which wraps any test with normal limiting distribution and adjusts critical values for rejection. This approach is not applicable to F-tests, and by extension efficiency tests, due to the normal distribution assumption. All other tests presented in section 4 can be applied in this setting. The fluctuation null hypothesis is rejected if it is rejected in any of the rolling windows. We illustrate the usefulness of these fluctuation tests using wage forecasts from a volatile environment.

Figure 5 first helps us to visualise the patterns emerging in the wage data. Notably, the time series of forecasts and the latest data outturns appear to show the MPR wage forecast systematically overpredicting wage growth up until the Covid period, with outturns below projections on average (ie negative biasedness in the forecast errors). Thereafter, it appears to systematically underpredict wage growth (ie positive biasedness in the errors). Figure 6 confirms this by showing the rolling estimated bias coefficient for wage growth forecasts from the Monetary Policy Report using a 16-quarter window at the 0, 4 and 8 quarter-ahead forecast horizons. The p-values from

fluctuation tests are also reported to assess the statistical significance of bias over time. This test confirms that wage forecasts significantly overpredicted outcomes at one- and two-years ahead before the pandemic, in part owing to weak productivity growth pre-Covid, but have significantly underpredicted them in the post-Covid period. Such a finding would typically lead the forecaster to investigate some of its potential drivers further. For example, in this case, Bank staff analysis has suggested:

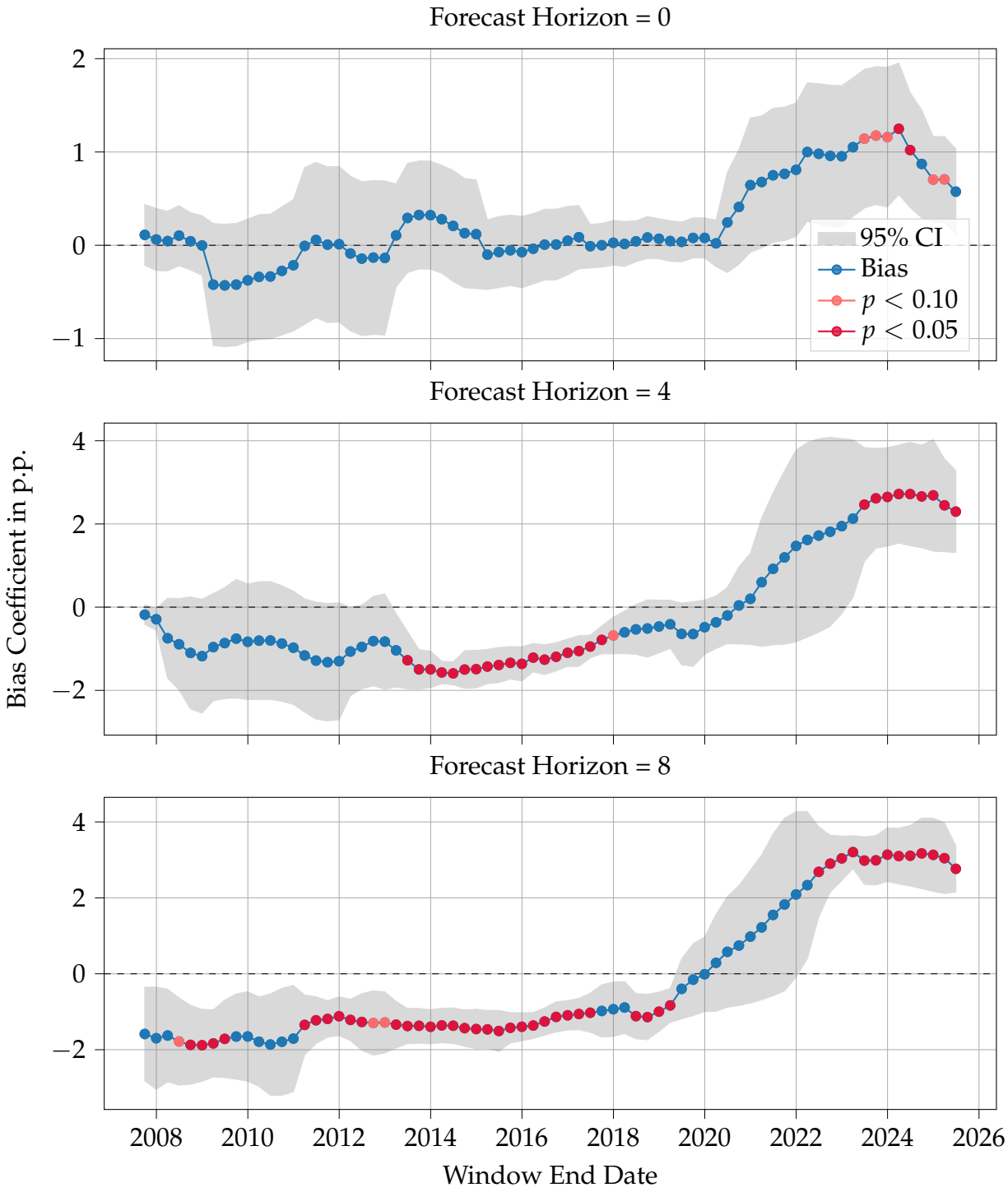
- The Bank's forecast underestimated post-Covid wage growth across 2020-2021 owing to compositional changes in the labour market, and the Job Retention Scheme (JRS).
- The Bank found it difficult to judge the extent of second-round effects on price and wage dynamics (Broadbent, 2022). The sharp rise in inflation over 2021-2022, owing to an external energy price shock, fed into higher inflation expectations domestically. This in turn fed into stronger than expected wage growth. The size of the external shock to inflation and the tightness in the labour market over 2021-2022 meant that estimates from historical episodes were unlikely to be a good guide to behaviour this time.
- In the post-Covid period, there also appeared to be evidence of structural change in the labour market that contributed to stronger wage growth. Wage bargainers may have exhibited strong real wage resistance following the external energy price shock (Pill, 2025). There also appeared to be evidence of reduced matching efficiency and search intensity, increased labour hoarding and recruitment difficulties, and declines in labour market participation (see Bank of England, 2025; Greene, 2024). These factors contributed towards a structurally tighter labour market, which helped to keep wage growth elevated. This is particularly evident over the 2021-2023 period, where positive bias worsens significantly, having previously exhibited a negative bias in the years prior (see Figure 6).

Wage growth is a key indicator of domestic inflationary pressure, making this evaluation highly relevant for MPC monetary policy decisions. The persistent underestimation of wage growth highlights that our wage forecast models struggled during the recent history of the labour market.

Where such challenges manifest, we employ multiple strategies to address them. One such example is that recent Monetary Policy Reports have incorporated scenarios illustrating how economic developments might unfold under a regime of higher wage and inflation persistence, enabling policymakers to factor these risks into their deliberations. In parallel to that we are also investing in enhanced modelling capabilities to improve forecasts of key variables following the recommendations of the Bernanke Review. For instance, how to better capture structural shifts or non-linearities in the labour market (Bank of England, 2026).

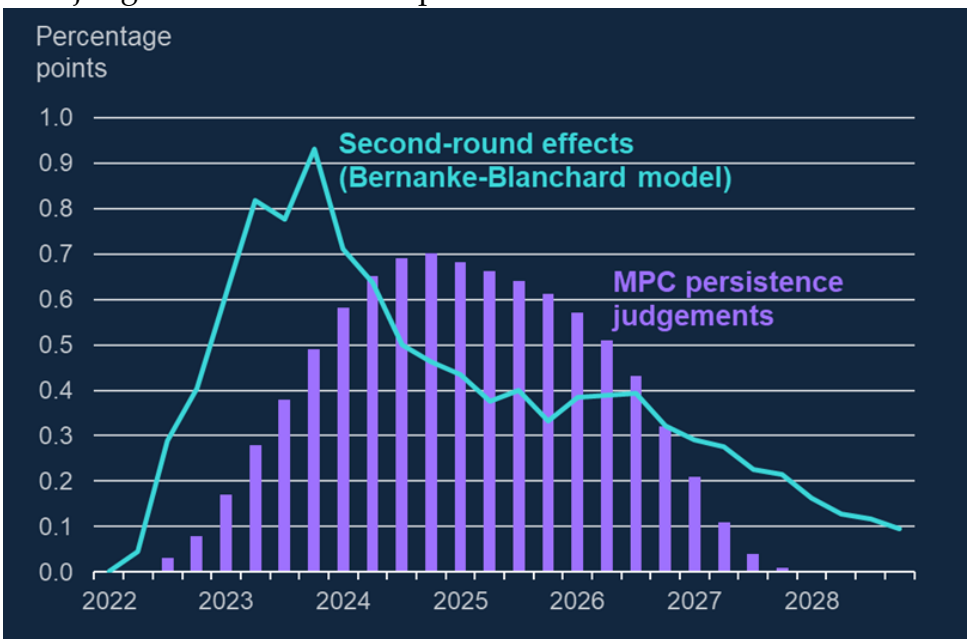
That said, we acknowledge that no macroeconomic model captures all features of the economy – every model is, after all, a simplification of reality. Therefore, utilising a range of different models and some degree of judgement will always be necessary. We also intend to more systematically record and evaluate the performance of judgements going forward, as part of our wider efforts to improve forecast evaluation. For example,

Figure 6: Rolling bias of forecasts for wage growth



This figure shows the rolling bias of MPR's forecasts of wage growth, calculated over a moving window of 16 observations. The fluctuation test evaluates bias across all windows jointly. P-values from the fluctuation test that are significant at the 10% and 5% levels are highlighted in red. These are more conservative than individual window tests when taken in isolation. The shaded region shows the 95% confidence interval computed separately for each window. This provides a useful guide to uncertainty around individual bias estimates, but cannot be used to interpret the fluctuation test, which evaluates bias across all windows jointly.

Figure 7: The Bernanke-Blanchard model estimate of second-round effects has informed MPC judgements on inflation persistence



This figure shows estimated second-round effects from the rise in inflation since 2021 in the Bernanke-Blanchard model (see Haskel et al., 2025) (aqua line) and MPC judgements for inflation persistence in the forecast (purple bars).

in 2023-2024, the MPC incorporated judgements in the wage and inflation forecasts to increase the persistence of these nominal variables, which have helped to reduce forecast errors. These judgements were informed both by the persistent forecast errors on wage growth, and by an adaptation of the Bernanke-Blanchard model to the UK context (see Bernanke and Blanchard, 2025; Haskel et al., 2025). This model suggested second-round effects added nearly 1 percentage point to inflation in late 2023 before beginning to unwind gradually (see Figure 7).

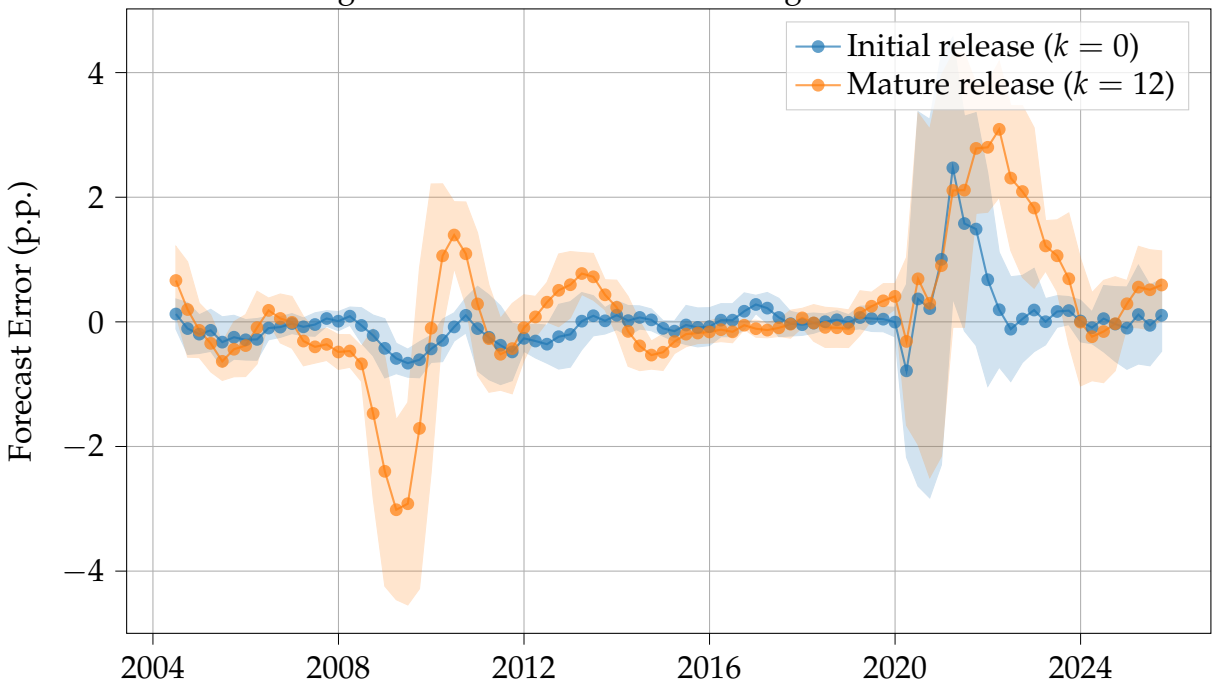
Working with incomplete data and the role of revisions

Third, data revisions pose another challenge for forecast evaluation in real time.

Macroeconomic data from national accounts are routinely subject to revisions as more data become available and methodologies improve. Labour market data are also subject to revision. As mentioned in section 4, we can define a 'mature' data release as one that is available 3 years after the fact ($k = 12$), providing a more accurate picture of the economy compared to the earliest available data within a given quarter ($k = 0$). The magnitude of revisions varies across series: while GDP can be subject to large revisions, headline CPI inflation data are not revised at all, although seasonal factors are updated over time when (in-house) seasonal adjustment is applied to the inflation series.

Bank staff must produce their analysis, and monetary policymakers set policy, on the basis of real-time information, but subsequent revisions can sometimes alter the data significantly. So it is important to also consider the role of this data uncertainty

Figure 8: Forecast errors for GDP growth



This figure shows MPR's 0-quarter ahead forecast errors of GDP growth using different outturn vintages. A moving average window of 4 quarters has been applied to smooth the series. The shaded area represents one standard deviation from the moving average.

when assessing if forecast performance (as measured against the mature data) could realistically have been improved on the basis of the information that was available at the time.

Revisions more specifically create two distinct issues for forecasters:

- The data available at the time of the forecast may be incomplete or inaccurate, leading to potential misjudgements about the current state of the economy.
- Revisions to the backdata can also directly change the jumping off point for forecasts and reduce the accuracy of year-on-year forecasts (which take those data as given) up to a year ahead.

Figure 8 illustrates the effect of evaluating GDP growth nowcasts using the earliest available data versus mature data published with a 3-year lag. Using the earliest available data ($k = 0$) for evaluation reveals that only the Covid period led to forecast errors above 1pp in absolute terms. However, using the mature estimate, there are multiple periods where the size of the absolute forecast error is above 1pp, most notably after the Global Financial Crisis and then after the Covid pandemic. This reflects how the interaction of the revisions process with large macroeconomic shocks can lead to larger errors, owing to the difficulty in pinning down the state of the economy in economic crises in real time.

5.2 Establishing causes and narratives

The tools presented above help us detect unusually large and persistent forecast errors and breaks in historical patterns. The unprecedented upside forecast error for inflation shown in Figure 4 and the shift in wage forecast bias documented in Figure 6, for instance, warrant further examination to understand their underlying causes.

This section aims to interrogate the drivers of recent forecast errors, through the lens of macroeconomic models. In particular, we seek to understand the role of identifiable factors, including developments in the conditioning paths used to produce MPR forecasts. Other factors account for the remainder of the error, and could include misspecification of the propagation of past shocks.

We present model-specific counterfactual exercises and historical decompositions that allow us to interrogate the sources of post-Covid forecast errors. As we explain further below, these exercises provide valuable insights but are inherently model-specific and uncertain. Each model embodies assumptions about the types and transmission of shocks, and can provide differing assessments of the drivers of past forecast errors.

5.2.1 Counterfactual forecasts

The conditional nature of the MPR forecasts mean that, with some assumptions, it is possible to separate out the role of deviations in conditioning variables from their projected paths, versus other factors, in accounting for forecast errors in key variables of interest. This is achieved by constructing 'counterfactual' exercises, taking a past forecast and asking the question 'what would this forecast have been if we had full information about the subsequent realisations of the conditioning paths in advance'.

Surprises in conditioning paths can be large, especially for energy prices in the recent past, and are by definition unanticipated due to the conditional nature of the MPR forecasts. They may account for a large proportion of recent forecast errors, all else equal. However, even after accounting for the role of conditioning path 'news', there will still be other factors that explain forecast errors. For instance, other identifiable sources of shocks may play a role; there may be shocks that are not captured by the model and therefore misidentified within it; or the model may be misspecified, such that the 'true' propagation of past shocks also contributes to errors. Whilst each of these explanations is plausible, inherent modelling and data uncertainty mean there remains an important role for judgement in understanding drivers of recent forecast errors and what signal to draw from these regarding the macroeconomic outlook.

These counterfactual exercises do, of course, have limitations themselves. The estimated impact of 'news' in conditioning paths is likely to be model-specific (as is the set of conditioning information) and dependent on the nature of the shocks hitting the economy ('state contingency'). Different models will thus provide differing explanations for the sources of past forecast errors. Moreover, the identification of the role of conditioning assumptions is subject to judgement about the reasons underlying their evolution. For example, monetary and fiscal policy will likely respond to the other shocks impacting the economy, but constraining these to follow conditioning paths removes that feedback mechanism in the MPR forecasts.

Given these limitations, future analysis could benefit from applying multiple modelling approaches to investigating the role of 'news' in forecast errors. Below, we present counterfactual forecasts for inflation and GDP using the Bank's core DSGE model, COMPASS (Albuquerque et al., 2025).

Model-specific counterfactual forecasts for CPI inflation and GDP

To investigate the role of news to conditioning assumptions in driving forecast errors of recent years, we build on a previous exercise presented in Albuquerque et al. (2025). Using the latest version of the COMPASS DSGE model, and taking account of the latest vintage of National Accounts data, we reconstruct the August 2021 MPR conditional forecasts for year-on-year CPI inflation and the level of GDP over the period 2021Q2-2024Q3. To do so, we specify a set of structural shocks in the model, as well as applying a Kalman filter, to rationalise and fully match actual observed outturns for the conditioning paths (based on data as of November 2025).

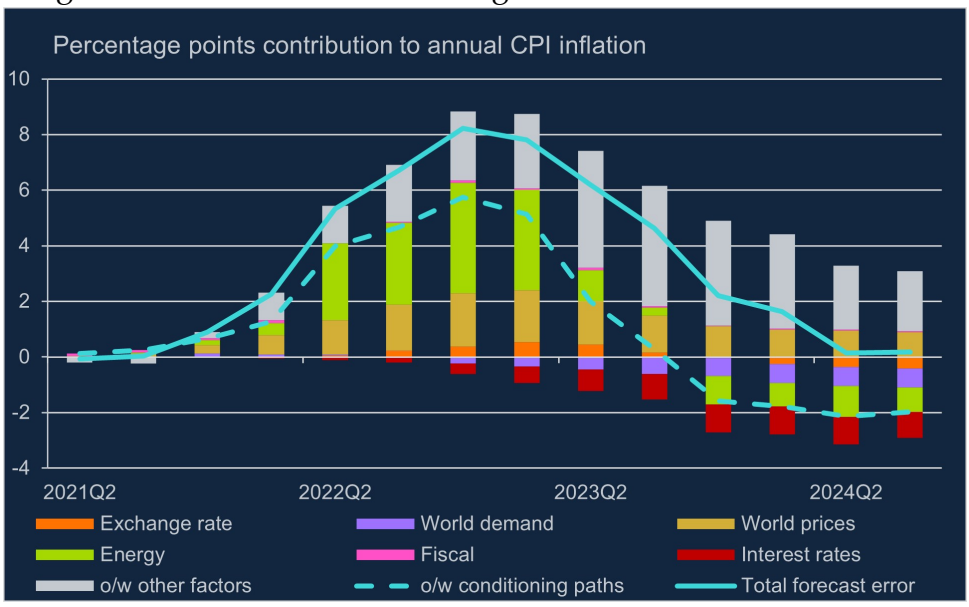
Our counterfactual forecasts cover the post-Covid period where energy prices, inflation and interest rates all peaked. COMPASS only includes a subset of conditioning variables used for MPR forecasts, namely Bank Rate, real government spending, the sterling effective exchange rate index, the direct contribution of energy prices to CPI inflation, and a set of projections for five world variables (capturing global demand, prices, and interest rates). Our counterfactual forecasts are produced using actual data values for these conditioning variables in the model, based on the data vintage included in the November 2025 MPR.

We show that in this DSGE model, relative to the August 2021 MPR forecast, news in the subset of the conditioning paths in COMPASS can account for more than two-thirds (5.7pp) of the 8.2pp upside forecast error for inflation at its peak in 2022Q4 (Figure 9). In particular, large upside news in energy prices (+3.9pp) and wider global prices (+1.9pp), alongside upside news from a depreciation in the sterling exchange rate (+0.4pp), are the largest contributors to the upside forecast error for inflation at its peak.

Conditioning path news also accounts for around four-fifths (1.6pp) of the 1.9% downside forecast error in the level of GDP at its trough in 2023Q4 (Figure 10). However, unlike for inflation, energy price news plays a smaller role in explaining GDP forecast errors, according to COMPASS. Instead, upside news in the market conditioning path for Bank Rate is estimated to have dragged on the level of GDP by up to 2.3% by 2024Q3 relative to the August 2021 MPR forecast. The combined drag from upside news in energy prices and the market conditioning path for Bank Rate more than accounts for lower GDP from 2022Q3 onwards, relative to the August 2021 MPR forecast, partially offset by upside news in real government spending, and with support from the depreciation in the exchange rate.

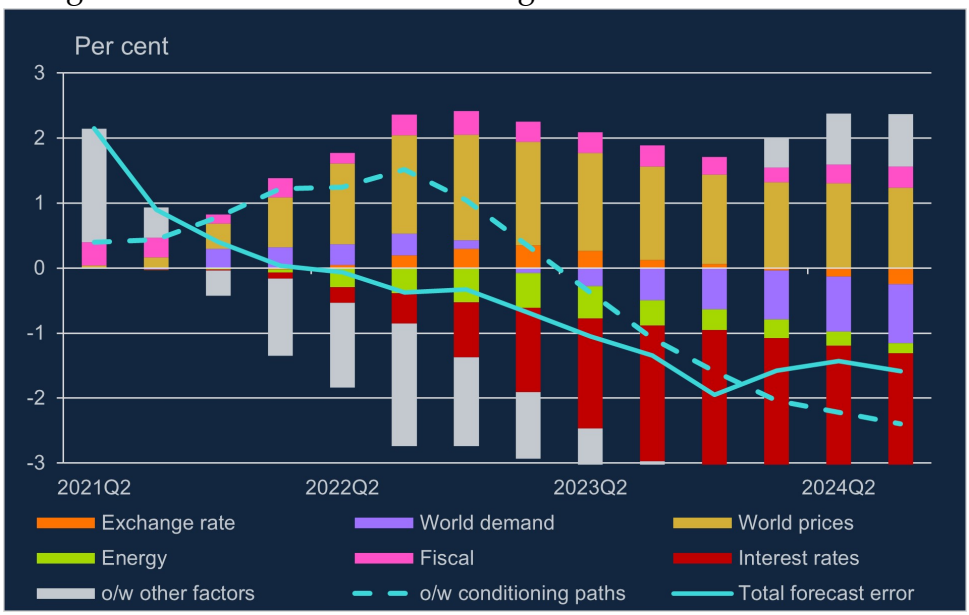
Conditioning paths news is a strong driver of recent forecast errors for inflation and GDP. However, factors beyond the conditioning paths also play an important role in driving forecast errors for both inflation and GDP. As highlighted above, these other factors can point to a range of issues in macroeconomic models, or potentially reflect exogenous shocks to variables outside the conditioning assumptions. Establishing a plausible explanation for forecast errors unexplained by conditioning paths news is

Figure 9: Model-based accounting for CPI inflation forecast error



This figure shows forecast errors for the August 2021 MPR projection of CPI inflation (solid aqua line), and the total contribution from conditioning-path news (dashed aqua line). Bars show percentage-point contributions from news in individual conditioning paths (solid) and other factors (grey). The dashed line equals the sum of the news contributions.

Figure 10: Model-based accounting for GDP level forecast error



This figure shows forecast errors for the August 2021 MPR projection of GDP level growth from the start of the forecast (solid aqua line), and the total contribution from conditioning-path news (dashed aqua line). Bars show percentage-point contributions from news in individual conditioning paths (solid) and other factors (grey). The dashed line equals the sum of the news contributions.

inherently judgement-based; we discuss some exercises and tools that can help inform these judgements in Section 5.2.2

5.2.2 Drawing on other models and research to test model narratives

It is often difficult to directly identify which endogenous features or parameters within a model may have contributed to past forecast errors, especially for large structural models like COMPASS. We can gain some insights from investigating any given model's ex-post interpretation of observed economic data (by decomposing historical data into contributions from different model-based shocks), and comparing this to interpretations offered by other models, as well as the wider academic literature. This comparison allows us to consider factors that may have contributed to previous endogenous forecast errors from the model, such as the types of shocks, differences in shock transmission, or structural economic features that are potentially not well captured in the forecasting model of interest.

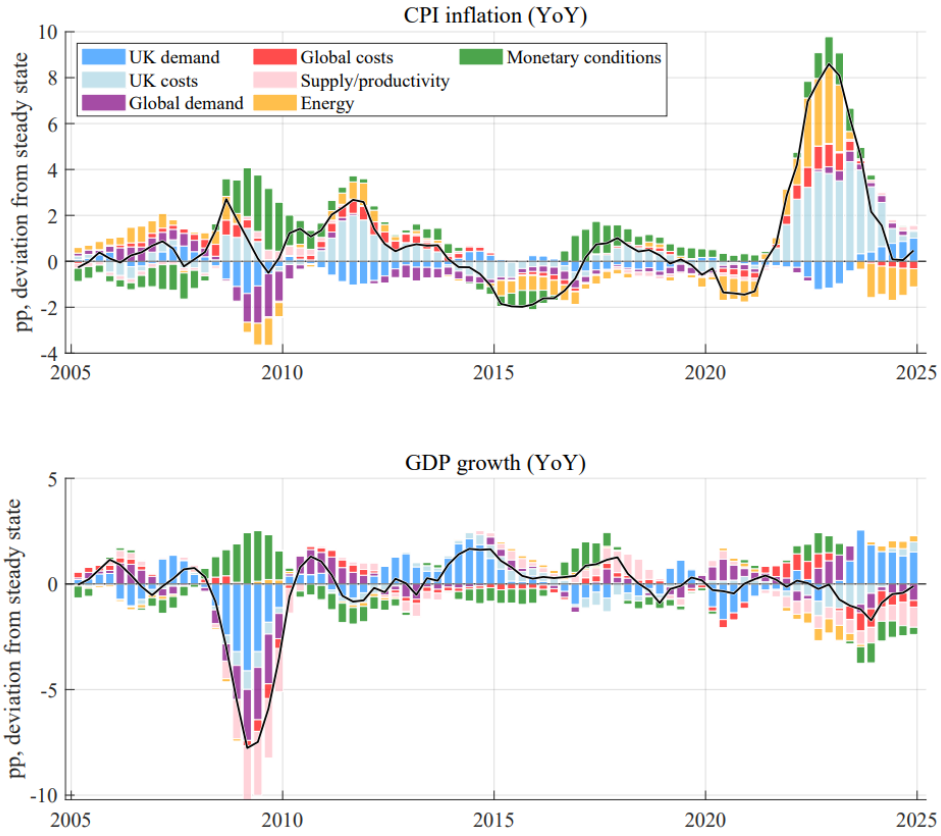
For example, COMPASS identifies robust UK demand conditions as a driver of GDP growth post-Covid, but estimates a limited role for demand in explaining inflation over the same period (Figure 11). This model-based interpretation differs from recent academic research, which suggests that strong demand was a key driver of post-pandemic inflation across advanced economies (Giannone and Primiceri, 2024; Bernanke and Blanchard, 2025). And suggests that further investigating the role of demand in the model (ie types and transmission of demand shocks, missing demand channels such as fiscal transfers, or demand-related structural model parameters) could perhaps help understand previous forecast errors for inflation.

Bank staff have recently developed an SVAR model for the UK that applies Bayesian methods to more clearly account for uncertainties in real-time economic interpretation and shock identification (Brignone and Piffer, 2025). The model uses forecast errors or revisions in successive time periods to make probabilistic statements about the types of shocks that may be hitting the economy at any given point in time. For example, it suggests that in 2024Q2 the UK economy was hit by a contractionary global demand shock with 74% probability, alongside an offsetting expansionary UK demand shock (Figure 12).

This probabilistic economic assessment addresses some of the challenges with model and data uncertainty mentioned earlier. However, any model is inherently a simplification of reality. This is why it is often valuable to use a range of specialised models to conduct more detailed assessments of the drivers of macroeconomic dynamics, to improve understanding of these mechanisms. For instance, to better understand post-Covid inflation dynamics, and as discussed earlier, Bank staff have recently applied Bernanke and Blanchard (2025) to UK data (Haskel et al., 2025). Machine learning techniques have also been used to explore potential non-linearities in price and wage setting after the pandemic (Buckmann et al., 2025), which may not be well captured in a linearised model like COMPASS (Albuquerque et al., 2025).

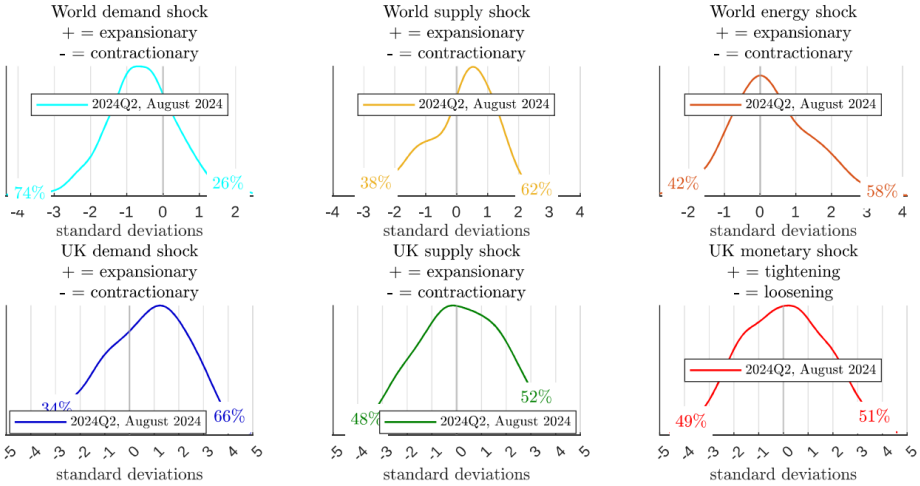
Understanding the drivers of past forecast errors inherently involves weighing competing steers from a range of macroeconomic models. Going forward, we intend to use the

Figure 11: Historical decompositions of CPI inflation and GDP growth from the COMPASS DSGE model (Albuquerque et al., 2025)



This figure shows the historical shock decompositions of Year-on-Year (YoY) CPI inflation and real GDP growth (in percentage points deviations from their steady states) using COMPASS (Albuquerque et al., 2025).

Figure 12: Distribution of shocks hitting the UK economy in 2024Q2 based on a SVAR model for the UK economy (Brignone and Piffer, 2025)



This figure shows the probabilistic distribution of shocks hitting the UK economy, identified by a UK SVAR model (Brignone and Piffer, 2025). Percentage point figures on the charts are the probability that the shock was negative (contractionary) or positive (expansionary).

enhanced toolkit and framework for forecast evaluation at the Bank to regularly identify and interrogate the sources of the Bank's forecast errors, informing key judgements and model development in real-time.

6 Conclusion

This paper has described and illustrated a range of techniques that underpin the Bank's refreshed approach to forecast evaluation. Through the development of a new forecast evaluation toolkit, we have enhanced our ability to conduct systematic assessment of forecast performance across multiple economic variables and time horizons.

Our new toolkit implements established techniques for measuring forecast accuracy, bias, and efficiency that are consistent with best practices in the literature. To automate these evaluation approaches, we have developed an open-source Python package that also facilitates diagnostics via a dashboard. We have demonstrated practical applications of this new framework by analysing Bank forecasts for GDP growth, CPI inflation, the unemployment rate, and wage growth, providing valuable insights into forecast performance.

The evaluation results highlight the importance of systematic forecast assessment in central banking. By comparing the Bank's forecasts against statistical benchmarks such as random walks and autoregressive processes, we can better understand the value added by expert judgement and sophisticated modelling. Combining this assessment with bias analysis allows us to identify potential shortcomings in our modelling suite. And the counterfactual exercises reveal the proportion of errors attributable to conditioning assumptions, thereby shedding light on the role of other factors.

Looking forward, this framework provides a foundation for ongoing forecast evaluation at the Bank of England. The modular design of our toolkit allows for the incorporation of additional evaluation metrics and techniques as the literature evolves. By releasing our toolkit as open-source software, we seek to enhance transparency around our processes and provide a shared platform that the wider community can contribute to. The systematic documentation of forecast performance over time will also contribute to institutional learning and support evidence-based improvements to the forecasting process.

References

Alati, Andrea, Martin Arazi, John Barrdear, Andrew Gimber, Elspeth Hughes, Lien Laureys, Simon Lloyd, Ozgen Ozturk, Jack Page, Kate Reinold, Eric Tong, Matthew Tong, and Matt Waldron (2025), "Tools for endogenous monetary policy analysis: optimal projections and instrument rules." Macro Technical Paper 4, Bank of England, URL https://www.bankofengland.co.uk/macro-technical-paper/2025/tools-for-endogenous-monetary-policy-analysis-optimal-projections-and-instrument-rules.

Albuquerque, Daniel, Jenny Chan, Derrick Kanngiesser, David Latto, Simon Lloyd, Sumer Singh, and Jan Žáček (2025), "Decompositions, forecasts and scenarios from an estimated DSGE model for the UK economy." Macro technical paper no. 1, Bank of England, URL https://www.bankofengland.co.uk/macro-technical-paper/2025/decompositions-forecasts-and-scenarios-from-an-estimated-dsge-model-for-the-uk-economy.

Anesti, Nikoleta, Simon Hayes, Andre Moreira, and James Tasker (2017), "Peering into the present: the Bank's approach to GDP nowcasting." *Bank of England Quarterly Bulletin*, URL https://www.bankofengland.co.uk/-/media/boe/files/quarterly-bulletin/2017/peering-into-the-present-the-banks-approach-to-gdp-nowcasting.pdf.

Argiri, Eleni, Stephen G. Hall, Angeliki Momtsia, Daphne Marina Papadopoulou, Ifigeneia Skotida, George S. Tavlas, and Yongli Wang (2024), "An evaluation of the inflation forecasting performance of the European Central Bank, the Federal Reserve, and the Bank of England." *Journal of Forecasting*, 43, 932–947, URL https://onlinelibrary.wiley.com/doi/epdf/10.1002/for.3063.

Bailey, Andrew (2025), "Monetary policy in uncertain times." URL https://www.bankofengland.co.uk/speech/2025/may/andrew-bailey-keynote-address-at-the-reykjavik-economic-conference-2025. Speech at the Reykjavík Economic Conference, Iceland.

Bank of England (1999), "Inflation Report: February 1999." Technical report, Bank of England, URL https://www.bankofengland.co.uk/-/media/boe/files/inflation-report/1999/february-1999.

Bank of England (2023), "Monetary Policy Report: August 2023." Technical report, Bank of England, URL https://www.bankofengland.co.uk/-/media/boe/files/monetary-policy-report/2023/august/monetary-policy-report-august-2023.pdf.

Bank of England (2024), "Monetary Policy Report: August 2024." Technical report, Bank of England, URL https://www.bankofengland.co.uk/-/media/boe/files/monetary-policy-report/2024/august/monetary-policy-report-august-2024.pdf.

Bank of England (2025), "Monetary Policy Report: November 2025." Technical report, Bank of England, URL https://www.bankofengland.co.uk/-/media/boe/files/monetary-policy-report/2025/november/monetary-policy-report-november-2025.pdf.

Bank of England (2026), "Forecast Evaluation Report: January 2026." Technical report, Bank of England, URL https://www.bankofengland.co.uk/paper/2026/forecast-evaluation-report-january-2026.

Bell, Venetia, Lai Wah Co, Sophie Stone, and Gavin Wallis (2014), "Nowcasting UK GDP growth." *Bank of England Quarterly Bulletin*, URL https://www.bankofengland.co.uk/quarterly-bulletin/2014/q1/nowcasting-uk-gdp-growth.

Bernanke, Ben and Olivier Blanchard (2025), "What Caused the US Pandemic-Era Inflation?" *American Economic Journal: Macroeconomics*, 17, 1–35, URL https://www.aeaweb.org/articles?id=10.1257/mac.20230195.

Bernanke, Ben S. (2024), "Forecasting for monetary policy making and communication at the Bank of England: a review." Technical report, Bank of England, URL https://www.bankofengland.co.uk/independent-evaluation-office/forecasting-for-monetary-policy-making-and-communication-at-the-bank-of-england-a-review.

Blanchard, Olivier J. and Daniel Leigh (2013), "Growth Forecast Errors and Fiscal Multipliers." *American Economic Review*, 103, 117–20, URL https://www.aeaweb.org/articles?id=10.1257/aer.103.3.117.

Boero, Gianna, Jeremy Smith, and Kenneth F. Wallis (2008), "Evaluating a Three-Dimensional Panel of Point Forecasts: The Bank of England Survey of External Forecasters." *International Journal of Forecasting*, 24, 354–367, URL https://www.sciencedirect.com/science/article/pii/S0169207008000526.

Bohm, Thomas and Marea Sing (2022), "Evaluating the Reserve Bank's forecasting performance." Bulletin, Reserve Bank of New Zealand, URL https://www.rbnz.govt.nz/hub/publications/bulletin/2022/rafimp-bulletin.

Bowe, Frida et al. (2023), "A SMARTer way to forecast." Staff Memo 7, Norges Bank, URL https://hdl.handle.net/11250/3061863.

Brignone, Davide and Michele Piffer (2025), "A Structural VAR Model for the UK Economy." Macro technical paper no. 3, Bank of England, URL https://www.bankofengland.co.uk/macro-technical-paper/2025/a-structural-var-model-for-the-uk-economy.

Broadbent, Ben (2022), "The Inflationary Consequences of Real Shocks." URL https://www.bankofengland.co.uk/speech/2022/october/ben-broadbent-speech-at-imperial-college-the-inflationary-consequences-of-real-shocks. Speech at Imperial College.

Buckmann, Marcus, Galina Potjagailo, and Philip Schnattinger (2025), "Blockwise Boosted Inflation: Non-linear determinants of inflation using machine learning." Staff Working Paper 1143, Bank of England, URL https://www.bankofengland.co.uk/working-paper/2025/blockwise-boosted-inflation-non-linear-determinants-of-inflation-using-machine-learning.

Burgess, Stephen, Emilio Fernandez-Corugedo, Charlotta Groth, Richard Harrison, Francesca Monti, Konstantinos Theodoridis, and Matt Waldron (2013), "The Bank of England's forecasting platform: COMPASS, MAPS, EASE and the suite of models." Staff working paper 471, Bank of England, URL https://www.bankofengland.co.uk/working-paper/2013/the-boes-forecasting-platform-compass-maps-ease-and-the-suite-of-models.

Cascaldi-Garcia, Danilo (2022), "Pandemic Priors: Forecasting and Policy Analysis with a Pandemic DSGE Model." *Finance and Economics Discussion Series*, 1–54, URL https://www.federalreserve.gov/econres/ifdp/files/ifdp1352.pdf.

Castle, Jennifer L., Jurgen A. Doornik, and David F. Hendry (2025), "Could the Bank of England have avoided mis-forecasting UK inflation during 2021-24?" *International Journal of Forecasting*, URL https://www.sciencedirect.com/science/article/pii/S0169207025000603.

Clements, Michael P. (2004), "Evaluating the Bank of England Density Forecasts of Inflation." *Economic Journal*, 114, 844–866, URL https://onlinelibrary.wiley.com/doi/10.1111/j.1468-0297.2004.00246.x.

Coibion, Olivier and Yuriy Gorodnichenko (2015), "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts." *American Economic Review*, 105, 2644–78, URL https://www.aeaweb.org/articles?id=10.1257/aer.20110306.

Coroneo, Laura (2025), "Forecasting for monetary policy." *International Journal of Forecasting*, URL https://www.sciencedirect.com/science/article/pii/S0169207025000457.

Daniell, Harvey and Andre Moreira (2023), "Forecasting near-term trends in the labour market." Bank underground, Bank of England, URL https://bankunderground.co.uk/2023/08/17/forecasting-near-term-trends-in-the-labour-market/.

Dhami, Pavandeep, Mike Goldby, Clare Macallan, Ben Nelson, Matt Tong, and Danny Walker (2025), "Monetary policymaking at the Bank of England in uncertain times." *Quarterly Bulletin*, URL https://www.bankofengland.co.uk/quarterly-bulletin/2025/2025/monetary-policymaking-at-the-bank-of-england-in-uncertain-times.

Diebold, Francis and Roberto Mariano (1995), "Comparing Predictive Accuracy." *Journal of Business & Economic Statistics*, 13, 253–63, URL https://EconPapers.repec.org/RePEc:bes:jnlbes:v:13:y:1995:i:3:p:253-63.

Esady, Vania and Mansi Mate (2025), "Monthly Trend Inflation Measurement at Sectoral Level.", URL https://www.lse.ac.uk/CFM/assets/pdf/CFM-Discussion-Papers-2025/CFMDP2025-12-Paper.pdf. CFM Discussion Paper No. CFMDP2025-12.

Giacomini, Raffaella and Barbara Rossi (2010), "Forecast comparisons in unstable environments." *International Journal of Forecasting*, URL https://onlinelibrary.wiley.com/doi/full/10.1002/jae.1177.

Giannone, Domenico, Michele Lenza, and Giorgio E. Primiceri (2015), "Prior Selection for Vector Autoregressions." *Review of Economics and Statistics*, 97, 436–451, URL https://direct.mit.edu/rest/article-abstract/97/2/436/58236/Prior-Selection-for-Vector-Autoregressions.

Giannone, Domenico and Giorgio Primiceri (2024), "The Drivers of Post-Pandemic Inflation." Working Paper 32859, National Bureau of Economic Research, URL http://www.nber.org/papers/w32859.

Greene, Megan (2024), "Two Puzzles: Recent UK Labour Market Dynamics." URL https://www.bankofengland.co.uk/speech/2024/may/megan-greene-speech-at-make-uk-the-current-state-of-britains-labour-market. Speech at Make UK.

Groen, Jan J. J., George Kapetanios, and Simon Price (2009), "A Real Time Evaluation of Bank of England Forecasts of Inflation and Growth." *International Journal of Forecasting*, 25, 74–80, URL https://www.sciencedirect.com/science/article/pii/S016920700800109X.

Haberis, Alex, Richard Harrison, Kate Reinold, and Matt Waldron (2025), "Monetary policymaking under uncertainty." Macro Technical Paper 5, Bank of England, URL https://www.bankofengland.co.uk/macro-technical-paper/2025/monetary-policymaking-under-uncertainty.

Harvey, David, Stephen Leybourne, and Paul Newbold (1997), "Testing the equality of prediction mean squared errors." *International Journal of Forecasting*, 13, 281–291, URL https://www.sciencedirect.com/science/article/pii/S0169207096007194.

Harvey, David I., Stephen J. Leybourne, and Emily J. Whitehouse (2017), "Forecast evaluation tests and negative long-run variance estimates in small samples." *International Journal of Forecasting*, 33, 833–847, URL https://www.sciencedirect.com/science/article/pii/S0169207017300559.

Haskel, Jonathan, Josh Martin, and Lennart Brandt (2025), "What explains recent UK inflation? An application of the Bernanke-Blanchard model." Working paper, Economic Statistics Centre of Excellence, URL https://www.escoe.ac.uk/publications/what-explains-recent-uk-inflation-an-application-of-the-bernanke-blanchard-model/.

Independent Evaluation Office (2015), "Evaluating forecast performance." Independent evaluation office, Bank of England, URL https://www.bankofengland.co.uk/-/media/boe/files/independent-evaluation-office/2015/evaluating-forecast-performance-november-2015.pdf.

Johansson, Jesper, Mårten Löf, Ard Den Reijer, Pär Stockhammer, and Anna Österberg (2023), "Evaluation of the Riksbank's forecasts." Riksbank studies, Sveriges Riksbank, URL https://www.riksbank.se/globalassets/media/rapporter/riksbanksstudie/engelska/2023/riksbank-study-evaluation-of-the-riksbanks-forecasts.pdf.

Kanngiesser, Derrick and Tim Willems (2024), "Forecast accuracy and efficiency at the Bank of England - and how errors can be leveraged to do better." Staff working paper, Bank of England, URL https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2024/forecast-accuracy-and-efficiency-at-boe-how-errors-leveraged-to-do-better.pdf.

Lane, Philip (2024), "The 2021-2022 inflation surges and the monetary policy response through the lens of macroeconomic models." URL https://www.ecb.europa.eu/press/key/date/2024/html/ecb.sp241118_1~2c31ddbaa8.en.html. Speech at the SUERF Marjolin Lecture hosted by the Banca dItalia.

Lombardelli, Clare (2024), "Managing the present, shaping the future." URL https://www.bankofengland.co.uk/speech/2024/november/clare-lombardelli-speech-at-the-3rd-boe-watchers-conference. Speech at the Bank of England Watchers Conference.

Marmol, Francesc (1995), "The Stationarity Conditions for an AR(2) Process and Schur's Theorem." *Econometric Theory*, 11, 1180–1182, URL http://www.jstor.org/stable/3532612.

Mincer, Jacob A. and Victor Zarnowitz (1969), *The Evaluation of Economic Forecasts*, none edition, volume None of *NBER Books*. National Bureau of Economic Research, Inc, URL https://www.nber.org/system/files/chapters/c1214/c1214.pdf.

Moreira, Andre (2025), "Nowcasting GDP at the Bank of England: A Staggered-Combination MIDAS Approach." Macro Technical Paper 2, Bank of England, URL https://www.bankofengland.co.uk/macro-technical-paper/2025/nowcasting-gdp-at-the-bank-of-england-a-staggered-combination-midas-approach.

Nordhaus, William D. (1987), "Forecasting Efficiency: Concepts and Applications." *The Review of Economics and Statistics*, 69, 667–674, URL http://www.jstor.org/stable/1935962.

Office for Budget Responsibility (2024), "Forecast Evaluation Report." Technical report, Office for Budget Responsibility, URL https://obr.uk/download/forecast-evaluation-report-october-2024/?tmstv=1757672966.

Office for National Statistics (2019), "Transformation of Gross Domestic Product in Blue Book 2019." URL https://www.ons.gov.uk/economy/nationalaccounts/uksectoraccounts/articles/nationalaccountsarticles/transformationofgrossdomesticproductinbluebook2019.

Pill, Huw (2025), "The Courage Not to Act." URL https://www.bankofengland.co.uk/-/media/boe/files/speech/2025/may/the-courage-not-to-act-remarks-by-huw-pill.pdf. Speech at Barclays.

Stockton, David (2012), "Review of the Monetary Policy Committee's Forecasting Capability." Technical report, Bank of England, URL https://www.bankofengland.co.uk/-/media/boe/files/news/2012/november/the-mpcs-forecasting-capability.pdf.

Waggoner, Daniel F. and Tao Zha (1999), "Conditional Forecasts in Dynamic Multivariate Models." *Review of Economics and Statistics*, 81, 639–651, URL https://www.jstor.org/stable/2646713.

Wallis, Kenneth F. (2004), "An Assessment of Bank of England and National Institute Inflation Forecast Uncertainties." *National Institute Economic Review*, 189, 64–71, URL https://www.cambridge.org/core/journals/national-institute-economic-review/article/abs/an-assessment-of-bank-of-england-and-national-institute-inflation-forecast-uncertainties/F0289026A55BA2E1B0F8B49DADCEBB48.

A Appendix

A.1 Data Science Approach to Forecast Evaluation

One of the core purposes behind creating the forecast evaluation toolkit is to support the annual production of FERs. More generally, the architecture and toolkit will help the Bank to embed forecast evaluation more systematically into our processes and modelling, which will help us to respond to recommendation two in Bernanke (2024).

As a result, we have adopted a long term, data science led approach by developing a dedicated Python package for forecast evaluation. To be as transparent as possible we have made this public at the following GitHub location:

https://github.com/bank-of-england/forecast_evaluation.

This package allows the results in section 4 of the paper to be reproduced, but more importantly provides a standardised and governed toolkit that any researcher can use to evaluate their own forecasts using the same methods, metrics, and economic principles applied to the Bank's core models.

The package provides a single, consistent layer for forecast evaluation across the organisation. For example, a researcher developing a new forecasting model, regardless of the language or platform they use, can submit their forecasts across K vintages to the package and obtain the same set of metrics and tests as every other researcher. This eliminates divergence in individual approaches and ensures comparability across models and over time.

Because the package is agnostic to both forecast sources and outturn data, it scales naturally. Researchers can apply the same evaluation methods to N models across N macroeconomic series and benchmark them against M other Bank models without writing new code or modifying the underlying workflow.

Developing the package also establishes the foundation for our future platform integration. As the Bank transitions its forecasting architecture onto Databricks, the package will be able to query outturn series directly via API, removing the need for users to manually curate or upload data. This will enable forecast evaluation at scale across any macro series available on the platform, provided a forecast is supplied.

More broadly, the package represents an important first step toward establishing consistency, reproducibility, and best practices across the directorate. Over time, economists will be able to contribute new metrics, tests, and methods, with updates automatically becoming available to all users. This shared framework strengthens governance, reduces duplication, and supports a more coherent and transparent forecast evaluation process across the Bank.

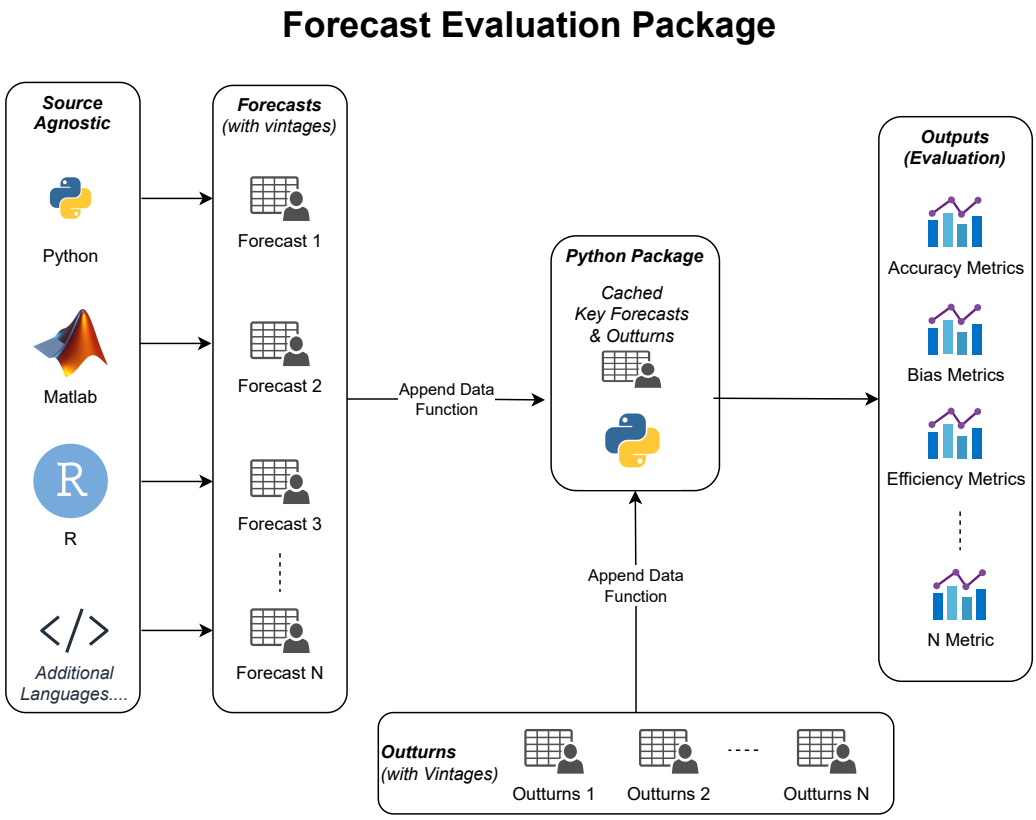


Figure 13: This figure shows the data architecture for the forecast evaluation toolkit.

Typical forecast evaluation workflow using the package is as follows (also illustrated in Figure 13):

- An Economist produces a new model any way they like for macro variable X
- A forecast is produced eg 12 quarters ahead forecast by vintage date for macro variable X
- The python package has an append function to read in the forecast
  - Any error or formatting issues are returned to the user to then resubmit
- The python package already caches the most important forecasts and outturns to enable comparison to the Bank's baselines models.
- If a macro variable outturn is more niche the user can also submit those to the package.
- Post forecast submission the python package can be used to produce an array of metrics, charts and dashboards.

A.2 Autoregressive Model Estimation

In this section, we provide more details on how the autoregressive models described in section 3.2.3 are estimated.

A.2.1 Model Specification

The AR(p) model with location-scale t-distributed errors is specified as:

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

where:

- y_t is the observed time series value at time t
- c is the constant term (intercept)
- ϕ_i are the autoregressive coefficients for $i = 1, 2, \ldots, p$
- ε_t are the error terms following a location-scale t-distribution

The error terms are distributed as:

$$\varepsilon_t \sim t(0, \sigma^2, \nu)$$

where σ is the scale parameter and ν is the degrees of freedom parameter.

The location-scale t-distribution is useful to use over the normal distribution for macroeconomic data as it has heavier tails to accommodate outliers, which reduces the impact of extreme observations in our parameter estimates.

A.2.2 Parameter Estimation

The parameters of the AR(p) model are estimated using Maximum Likelihood Estimation (MLE). The log-likelihood function for the AR(p) model with t-distributed errors is:

$$\mathcal{L}(\theta) = \sum_{t=p+1}^{T} \log f(\varepsilon_t; \sigma, \nu)$$

where $\theta = \{c, \phi_1, \ldots, \phi_p, \sigma, \nu\}$ is the parameter vector, T is the sample size and f is the probability density function of the location-scale t-distribution.

A.2.3 Optimisation Procedure

The MLE optimisation follows these steps:

The limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm with bound constraints (L-BFGS-B) is used for optimisation, which is a suitable method to use when we want to impose bound constraints on parameters such as the scale parameter and degrees of freedom.

Algorithm 1 MLE Parameter Estimation

1: Initialize parameters using OLS estimates
2: Set parameter bounds: $\sigma > 0.001$, $2 < \nu \leq 30$
3: Apply stationarity transformations to AR coefficients
4: Optimise negative log-likelihood using L-BFGS-B method
5: Calculate BIC for model selection

A.2.4 Initial Parameter Estimation

Initial parameter estimates are obtained via Ordinary Least Squares (OLS):

- AR coefficients and constant: OLS regression on lagged variables
- Scale parameter: Standard deviation of OLS residuals
- Degrees of freedom: Conservative initial value of 5.0

A.2.5 Stationarity Constraints

The stationary condition for AR(p)-models is that the roots of the characteristic polynomial must lie outside the unit circle. The characteristic polynomial for the AR(p) model is given by:

$$P(z) = 1 - \sum_{i=1}^{p} \phi_i z^i$$

The roots of this polynomial must satisfy:

$$|z_i| > 1 \quad \text{for all roots } z_i$$

This ensures that the time series is stationary and does not exhibit explosive behaviour.

For AR(1) and AR(2) models, we can enforce stationarity through transformations of the parameters during optimisation. The transformations are designed to ensure that the estimated coefficients remain within the bounds that guarantee stationarity. Moreover, the transformations we choose are smooth, which allows us to use gradient-based optimisation methods such as 'L-BFGS-B' for the MLE estimation of the parameters.

AR(1) Stationarity

For AR(1) models, stationarity requires:

$$|\phi_1| < 1$$

This is enforced using the transformation:

$$\phi_1 = \tanh(\phi_1^{raw}) \times 0.99$$

where ϕ_1^{raw} is the unconstrained parameter from optimisation. The range of tanh is $(-1, 1)$, and we multiply by 0.99 so that the transformed value does not get too close to the boundary.

AR(2) Stationarity

Following Marmol (1995), an alternative characterisation for stationarity for AR(2) models, is that the coefficients lie within the triangular region defined by:

$$\begin{aligned} \phi_1 + \phi_2 &< 1 \\ \phi_2 - \phi_1 &< 1 \\ |\phi_2| &< 1 \end{aligned}$$

The transformation procedure we apply is:

1. Transform ϕ_2 by:
$$\phi_2 = \tanh(\phi_2^{raw}) \times 0.99$$
2. Transform ϕ_1 by:
$$\phi_1 = (1 - \phi_2) - \frac{\phi_2 - 1}{1 + e^{-\phi_1^{raw}}}$$

The first transformation ensures that ϕ_2 remains within the bounds of $(-1, 1)$, while the second transformation ensures that ϕ_1 satisfies the triangular region constraints.

A.2.6 Model Selection

The optimal lag order p is selected using the Bayesian Information Criterion (BIC):

$$\text{BIC}(p) = -2\mathcal{L}(\hat{\theta}_p) + (p + 3)\ln(T - p)$$

where $(p + 3)$ accounts for the AR coefficients, constant, scale, and degrees of freedom parameters. We restrict the maximum lag order to $p_{\max} = 2$ to keep the model simple, but also because the transformation procedure for stationarity becomes more complex for higher orders. We evaluate the BIC for $p = 1$ and $p = 2$ and select the model with the lowest BIC value.

Algorithm 2 Lag Order Selection

```
Set maximum possible lag p_max = 2
Initialize best_BIC = ∞, optimal_lag = 1
for p = 1 to p_max do
    Fit AR(p) model using MLE
    if model converged AND BIC(p) < best_BIC then
        best_BIC ← BIC(p)
        optimal_lag ← p
    end if
end for
Return optimal_lag
```

A.2.7 Forecasting Procedure

Out-of-sample forecasts are generated using the fitted AR(p) model. The forecasts are produced by iteratively applying the AR(p) model equation:

$$\hat{y}_{T+h} = c + \sum_{i=1}^{p} \phi_i \hat{y}_{T+h-i}$$

for $h = 0, 1, \ldots, 12$ quarters ahead, where:

- $\hat{y}_{T+h-i} = y_{T+h-i}$ for $h - i < 0$ (historical values)
- $\hat{y}_{T+h-i}$ are previously forecasted values for $h - i \geq 0$

A.2.8 Data Processing and Implementation

The methodology applies the following data filters:

- Data from July 1997 onwards (post-BoE independence)
- Only variables in quarter-on-quarter (QoQ) change

We don't apply any additional filters such as using dummy variables to filter out extreme observations such as those from Covid-19, Brexit or the 2008 financial crisis. Instead, we rely on the robustness of using the location-scale t-distribution to handle outliers in the data.

A.3 Determining Standard Errors of the Wald Ratio Estimates

In this section, we provide details on how the standard errors of the Wald ratios in the Blanchard-Leigh regressions for strong efficiency in section 4.4.2 are determined.

The delta method applied to the Wald ratio $\omega(\beta, \delta) = \frac{\beta}{\delta}$, yields

$$\mathrm{SE}(\omega(\beta,\delta)) = \sqrt{\left(\frac{\partial\omega(\beta,\delta)}{\partial\beta}\right)^2 \mathrm{Var}(\beta) + \left(\frac{\partial\omega(\beta,\delta)}{\partial\delta}\right)^2 \mathrm{Var}(\delta) + 2\left(\frac{\partial\omega(\beta,\delta)}{\partial\beta}\right)\left(\frac{\partial\omega(\beta,\delta)}{\partial\delta}\right)\mathrm{Cov}(\beta,\delta)}$$

where $\mathrm{Var}(\beta)$ and $\mathrm{Var}(\delta)$ are the variances and $\mathrm{Cov}(\beta, \delta)$ the covariance of the estimates β and δ, respectively.

The partial derivatives are given by

$$\frac{\partial\omega(\beta,\delta)}{\partial\beta} = \frac{1}{\delta}, \quad \text{and} \quad \frac{\partial\omega(\beta,\delta)}{\partial\delta} = -\frac{\beta}{\delta^2}.$$

Therefore, the standard error of the Wald Ratio can be simplified to:

$$\mathrm{SE}(\omega(\beta,\delta)) = \sqrt{\frac{1}{\delta^2}\mathrm{Var}(\beta) + \frac{\beta^2}{\delta^4}\mathrm{Var}(\delta) - 2\frac{\beta}{\delta^3}\mathrm{Cov}(\beta,\delta)}.$$