International Journal of Forecasting 35 (2019) 100–120

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting the UK economy: Alternative forecasting methodologies and the role of off-model information ☆

Lena Boneva [a], Nicholas Fawcett [b], Riccardo M. Masolo [c,*], Matt Waldron [d]

[a] Bank of England and CEPR, United Kingdom
[b] Goldman Sachs International, United Kingdom
[c] Bank of England and CfM, United Kingdom
[d] Bank of England, United Kingdom

ARTICLE INFO

Keywords:
DSGE models
Forecasting
Financial crisis

ABSTRACT

How did DSGE model forecasts perform before, during and after the financial crisis, and what type of off-model information can improve the forecast accuracy? We tackle these questions by assessing the real-time forecast performance of a large DSGE model relative to statistical and judgmental benchmarks over the period from 2000 to 2013. The forecasting performances of all methods deteriorate substantially following the financial crisis. That is particularly evident for the DSGE model's GDP forecasts, but augmenting the model with a measure of survey expectations made its GDP forecasts more accurate, which supports the idea that timely off-model information is particularly useful in times of financial distress.

1. Introduction

Monetary policy relies on both accurate forecasts and an understanding of the transmission of shocks to the macroeconomy. While purely statistical models are usually hard to beat when considering the forecast accuracy, structural models can provide an economic interpretation of the economic outlook. Thus, central banks attach a premium to structural models producing accurate forecasts.

This paper shows how integrating off-model information into a quantitative DSGE can improve its forecasting performance materially, especially when large deviations from trend occur, as was the case in the Great Recession. In particular, we condition our DSGE model forecasts on inflation nowcasts, market-based expectations of future policy rates and survey-based measures of the expected output growth. Throughout, the off-model information is applied in such a way as to retain the underlying structure of the model.

In addition, we also benchmark the performance of the DSGE model against statistical and judgmental forecasts. We make our experiment more concrete by comparing an estimated DSGE model used at the Bank of England, known as COMPASS (Burgess et al., 2013), to a Suite of Statistical Models (Kapetanios, Labhard, & Price, 2008), which combines forecasts from several econometric models, and to the Monetary Policy Committee's judgmental forecasts, published in the quarterly *Inflation Report*.

We evaluate the forecast accuracy of our model-based forecasts by re-estimating both COMPASS and the Statistical Suite in each quarter between 2000Q1 and 2013Q1 using real-time data that reflect only information that was available at the time when the MPC's forecasts were produced. This real-time forecasting exercise makes use of an archive of MPC forecasts for inflation and GDP growth and

☆ *Any views expressed are solely those of the authors and so cannot be taken to represent those of the Bank of England or members of the Monetary Policy Committee, Financial Policy Committee, Prudential Regulation Authority Board or Goldman Sachs International. Nicholas Fawcett contributed to this article while employed by the Bank of England.*

* Corresponding author.
E-mail addresses: Lena.Boneva@bankofengland.co.uk (L. Boneva), Riccardo.Masolo@bankofengland.co.uk (R.M. Masolo), Matthew.Waldron@bankofengland.co.uk (M. Waldron).

https://doi.org/10.1016/j.ijforecast.2018.06.005
0169-2070/

the corresponding data vintages for a range of macroeconomic variables that has been being stored by the Bank of England since 1997.[1] Each data vintage is used to produce point and density forecasts from COMPASS and the Statistical Suite that are evaluated using a variety of statistical methods.

The MPC's forecasts are presented as so-called 'fan charts', which, as stated in the *Inflation Report*, represent "the MPC's best collective judgment about the most likely paths for inflation and output and the uncertainties around those central projections". These fan charts provide us with judgmental point and density forecasts, the accuracy of which we can assess in the same way as the models' forecasts.

While it is impossible to isolate a single model that outperforms the rest at all horizons and historical periods, it is possible to identify when a certain model performs particularly well, with the differences being statistically significant in some cases, despite the relatively short sample. At horizons of up to about a year, the *Inflation Report* forecasts are the most accurate for both GDP growth and inflation. However, at longer horizons, COMPASS and the Statistical Suite outperform the *Inflation Report* forecasts. This pattern is observed for both point and density forecasts.

The evaluation period includes both the run-up to the financial crisis and its aftermath. This was a period of huge macroeconomic volatility, including the deepest UK recession in the post-war period, which followed an unusually tranquil phase of low and stable inflation, and consistently positive GDP growth. Against this backdrop, it is not surprising that the forecasts proved to be inaccurate beyond one or two quarters ahead. This is true of all of the forecast methods that we consider in this paper, with none predicting the depth of the recession correctly at the onset of the financial crisis, or the combination of sluggish growth and resilient inflation that followed. This echoes the experience of other central banks and the wider literature (e.g. Del Negro & Schorfheide, 2012).

Although all of the forecast methods perform badly following the financial crisis, the comparative drop in accuracy of COMPASS's GDP forecasts is particularly marked. This is due to the quick trend-reversion mechanism that is built into most DSGEs, coupled with a lack of fast-moving forward-looking variables in the set of observables, which delays the model's "perception" of the severity of the crisis. Thus, an interesting question is whether different types of off-model information can improve the forecast performance in times of financial crisis and beyond. For example, we document that a variant of our DSGE model in which we augmented the set of observable variables with short-run expectations of GDP growth performs better for forecasting GDP growth.[2] In contrast to the model's other observables, survey expectations provide the model with timely off-model information that is particularly useful in times of financial distress. We also explore other types of off-model information, such as market-based forecasts of the policy rate and inflation nowcasts.[3]

A general result is that the effect on the forecast accuracy of broadening a model's information set depends on the way in which that information is applied. For example, when conditioning paths for the policy rate are imposed using a monetary policy shock, the forecast performance is worse than that of the model without additional information. However, applying the same off-model information using the full range of shocks improves the forecast accuracy.[4] The different methods of applying the same set of off-model information reflect alternative interpretations of the underlying reason for any difference between market expectations of the policy rate and the forecast from the model. For example, the interpretation of conditioning the forecast on market expectations of the policy rate using the monetary policy shock is that market participants would attribute any difference between their expectations and the plain model forecasts to non-systematic changes in the policy rate. In contrast, imposing the conditioning information using all shocks reflects the idea that market participants' expectations can differ from the model-implied forecasts along any dimension, e.g., the level of potential supply or the level of cost pressures.

The literature assessing the forecasting performances of DSGE models is extensive. For the most part (Del Negro & Schorfheide, 2012; Edge & Gürkaynak, 2011; Gürkaynak, Kisacikoglu, & Rossi, 2013; Wolters, 2013), it has investigated forecasts from small-scale models, similar to Christiano, Eichenbaum, and Evans (2005) and Smets and Wouters (2003). However, there has also been some work on central-bank models (which tend to be markedly larger), including the Swedish Riksbank (Iversen, Laseen, Lundvall, & Söderström, 2016), the ECB (Christoffel, Coenen, & Warne, 2010), and the Federal Reserve Board (Edge, Kiley, & Laforte, 2010). These studies all differ in the comparator forecasts that are used for evaluating the DSGE model and in the construction of the data for estimation and evaluation. For example, Edge and Gürkaynak (2011) use real-time data in their recursive re-estimation, whereas Iversen et al. (2016) do not. In addition, there is a substantial degree of variation in the statistical methods that are used for evaluating the forecasting performances of the DSGE models. While all papers report root mean squared forecast errors, only some perform statistical tests for assessing whether the forecasting performances of alternative models are significantly different (Edge & Gürkaynak, 2011; Edge et al., 2010; Gürkaynak et al., 2013). We formally test whether the accuracy of alternative forecasts is significantly different, and we also

[1] The full archive of real-time datasets used is available at: https://www.bankofengland.co.uk/working-paper/2015/evaluating-uk-point-and-density-forecasts-from-an-estimated-dgse-model-the-role-of-off-model.

[2] Note that we do not overwrite the GDP growth rate implied by our quarterly model in any specific quarter, but only restrict the average growth rate over four quarters to equal our measure of survey expectations.

[3] There are of course many other possible ways of augmenting DSGE models with more timely information or information that comes from less structural models. For example, Del Negro and Schorfheide (2004) propose a DSGE-VAR, while Giannone, Monti, and Reichlin (2016) and Cervena and Schneider (2014) add monthly data releases to a quarterly DSGE model.

[4] As does the exponential tilting method of Robertson, Tallman, and Whiteman (2005).

investigate instabilities in forecast performance, for both point and density forecasts.

However, compared to that literature, there is little work that assesses the effect of off-model information on model-based forecasts. One notable exception is the paper by Del Negro and Schorfheide (2012).[5] Unlike Del Negro and Schorfheide, we introduce off-model information without changing the transmission mechanism of the model — exploiting the fact that the COMPASS has more shocks than observables — which enables us to study the effects of survey-based expectations in the same model economy. Also, our experiment uses survey data not to inform long-run growth, as Del Negro and Schorfheide's did, but rather as a means of incorporating off-model information about the conjuncture.[6] Unlike Del Negro and Schorfheide (2012), our results suggest that survey expectations do play a significant role in altering the forecast accuracy. Given the difference in the horizon of the survey data, this result can be interpreted as being more in line with what Del Negro and Schorfheide (2012) found when they added a measure of credit spreads to the set of observables, in that it suggests that conditioning on timely conjunctural information can produce significant forecast accuracy gains.

The remainder of this paper is organized as follows. Section 2 introduces COMPASS, the Statistical Suite and the *Inflation Report* forecasts. Section 3 describes the data we use for estimation and forecast evaluation. Section 4 discusses the statistical methods used. Section 5 reports results on the accuracy of alternative point and density forecasts, and Section 6 investigates how off-model information can improve the DSGE model forecasts. Section 7 concludes.

2. Forecasting models and the *Inflation Report*

Each quarter, Bank of England staff produce an *Inflation Report* on behalf of the Monetary Policy Committee that contains, among other things, the MPC's best collective judgment density forecasts for inflation and GDP growth. In 2011, the Bank of England introduced a new forecasting platform to assist with the production of these *Inflation Report* forecasts (Burgess et al., 2013). This paper evaluates the performances of the forecasts from two key models that form part of that platform, relative to those of the *Inflation Report* forecasts themselves, focusing on the evaluation of an estimated DSGE model.

The new forecasting platform was designed with the aim of optimally supporting the process of producing the MPC's forecasts. Given the judgmental nature of those forecasts, this process places a premium on the discussion of the economics of the forecast, rather than on a desire to maximize the forecast accuracies of all of the models used. To reflect this, the new platform is built around a prototypical DSGE model, the *Central Organizing Model for Projection Analysis and Scenario Simulation* (COMPASS), which is used by the staff as an organizing device in the construction of the MPC's judgmental forecasts. Since COMPASS is built on economic theory, it can provide a structural interpretation and narrative around the MPC's forecasts, which can make useful contributions to staff discussions with the MPC. Of course, all models have relative strengths and weaknesses, and COMPASS is no exception, so a key design feature of the forecasting platform is that it recognizes the importance of a suite of models. The suite contains a large number of different models of varying types and classes, each with a different purpose in mind. Given the relative strengths and weaknesses of COMPASS, a particularly important set of models in that class is the "Statistical Suite of forecasting models", which have been designed explicitly with the forecasting performance in mind. The rest of this section provides some brief background on COMPASS, the Statistical Suite and the *Inflation Report* forecasts.

2.1. COMPASS

COMPASS is an open-economy New Keynesian DSGE model that is built on the tradition of Smets and Wouters (2003) and Christiano et al. (2005), and has similarities to antecedents at other central banks. The model economy – represented diagrammatically in Fig. 1 – is populated by households, firms, a central bank, a government, and an exogenous rest-of-the-world economy. The decisions made by each of these sectors result in laws of motion for key macroeconomic variables that are derived from assumptions about preferences, technologies and constraints. Like similar models used in other central banks, COMPASS incorporates nominal rigidities in price and wage settings, as well as a range of real rigidities, including habit formation in consumption and investment adjustment costs. For a detailed derivation and description of the model equations, we refer readers to the work of Burgess et al. (2013).

COMPASS is estimated using Bayesian maximum likelihood methods on UK data for 15 macroeconomic time series[7] with 18 shocks. One of those shocks is a permanent labor-augmenting productivity shock, which shifts the stochastic trend of the model, reflecting a statistical assumption that GDP and the expenditure components of GDP are integrated of order one and cointegrated with each other. The parameters of the model are divided into two groups. The first group of parameters are calibrated. This group predominantly comprises parameters that govern the steady state and trend properties of the model (e.g. share of consumption in output, trend growth rate of productivity, etc.). The second group of parameters are estimated using Bayesian maximum likelihood and mainly include parameters that govern the model's dynamics

[5] Benes, Binning, and Lees (2008) present an extensive analysis of the effects of conditioning on off model profiles for certain variables but they focus primarily on a plausibility index for the off-model judgements rather than a systematic forecast performance evaluation. Iversen et al. (2016) also consider conditional forecasts, but they do not include survey measures, while we can exploit survey-based GDP expectation that are published in the *Inflation Report* and we also experiment with alternative ways to impose the series we condition our forecast on.

[6] Long-run growth expectations are not available in the survey we use.

[7] These are: GDP, consumption, business investment, government expenditure, exports, imports, the export deflator, the import deflator, an index of average weekly earnings, seasonally adjusted consumer price inflation, the Bank Rate, the sterling effective exchange rate, total hours worked, an in-house measure of world trade, and an in-house measure of world export prices.

(e.g. cost of wage adjustment, degree of habit formation in consumption, etc.). As part of the real-time forecasting exercise described in Section 3, we re-estimate COMPASS recursively using real-time data following the same strategy as was described in detail in Section 4.3 of Burgess et al. (2013).[8]

In doing so, we extend the estimation sample beyond 2007Q4, which was the end of the estimation sample used by Burgess et al. (2013). This poses additional challenges that they did not tackle. In particular, the MPC cut the bank rate to 0.5% and implemented a quantitative easing (QE) program. Given the fact that COMPASS does not articulate a role for QE, and the practical difficulties of properly taking into account an occasionally binding constraint on interest rates like the zero lower bound, we use a 'shadow' measure of the policy rate as the observable policy rate rather than the bank rate as per Burgess et al. (2013). This shadow measure augments the bank rate to include an in-house estimate of the effect of QE.[9]

Section 5 evaluates COMPASS against the judgmental and statistical benchmarks that are described in the remainder of this section. Section 6 then investigates whether the forecast performance of COMPASS can be improved by augmenting it with off-model information such as market-based measures of future policy rates, for example.

2.2. Statistical suite

The bank's Suite of Statistical Forecasting Models offers a benchmark against which the COMPASS forecasts can be compared. This suite comprises a range of statistical models that do not have a particular economic structure in the way that COMPASS does; instead, the models form a set of 'reduced-form' relationships between macroeconomic variables.

The suite includes both univariate and multivariate forecasting methods. The set of univariate models covers a random walk, and autoregressive and smooth-transition models. Among the multivariate models are VARs, Bayesian VARs and data-rich methods such as factor models. The models have fixed parameters, so that they do not vary over time, but are all estimated over a seven-year rolling window. This provides a degree of robustness to structural change. The forecasts obtained from the individual models are then averaged to produce combined point and density forecasts, using weights based on each model's predictive likelihood. Further details of the models are provided in Appendix.

In addition to presentational advantages, there are also good practical grounds for model averaging. By combining many misspecified models, each of which incorporates information from different variables, model averaging usually outperforms forecasts from individual models (Aiolfi, Capistran, & Timmermann, 2010). Another important reason for poor forecast performance can be the presence of structural breaks (Clements & Hendry, 1998). Averaging can be robust to such breaks because not all models necessarily fail at the same time.

2.3. Inflation Report

Ever since monetary policy independence in 1997, the Bank's quarterly *Inflation Report* has been communicating the MPC's assessment of the economic outlook. As part of that assessment, the MPC produce 'fan charts', which represent their best collective judgment of the most likely paths for inflation and GDP growth, and the uncertainty surrounding them.[10] These fan charts provide us with a sequence of real-time judgmental point and density forecasts for inflation and GDP growth against which we can compare the model-based forecasts.

Two points stand out when making this comparison. First, by convention, the *Inflation Report* forecasts take the paths of several variables as given, including a particular path for monetary policy.[11] The statistical techniques that we use are valid under the assumption that the forecasts we evaluate are 'primitives', so this issue should be borne in mind when comparing the *Inflation Report* forecasts with the unconditional model-based forecasts. Secondly, there is no mechanical link between the model forecasts described in this paper (or any other model forecast) and the *Inflation Report* projections. The models are used by the staff to aid the deliberations of the MPC, but the final published forecast represents the Committee's best collective judgment. Thus, the forecast performance of one does not determine that of the other.

3. A real-time dataset for estimation and forecast evaluation

This section describes the real-time dataset that is used for estimating the models and evaluating the forecast accuracy. Throughout, our objective is to evaluate the models' forecast accuracy using only information and data that were available at the time when the corresponding *Inflation Report* forecast was made.[12] In addition, in order to

[8] This includes using the same prior distributions and the same estimation sample beginning in 1993Q1, with data from 1987Q2 to 1992Q4 being used as a training sample for the Kalman filter.

[9] The shadow rate is derived by computing a sequence of unanticipated monetary policy shocks to match the time series for the estimated effect of QE on GDP using estimates from Joyce, Tong, and Woods (2011); see also Section 8.4 of Burgess et al. (2013). The assumption that underlies this approach is that QE is a close substitute as a monetary policy instrument to the bank rate such that the zero lower bound was not an effective constraint on monetary policy over the period in question.

[10] Technically speaking, the density forecasts are constructed from two-part normal distributions parameterized with the mode, variance and skew.

[11] Since August 2004, the headline *Inflation Report* projections have been conditioned on market expectations for interest rates out to a three-year horizon. Prior to that, the headline projections were conditioned on constant rates, with market-rate-conditioned projections published alongside.

[12] There are limits on the extent to which our exercise can be regarded as truly 'real time'. For example, although the dataset that we use for recursive model estimation is a real-time dataset, the models themselves are not real time. The content and implementation of both COMPASS, which was introduced in 2011, and the Statistical Suite, which was introduced in 2005, have been influenced by developments in economic theory and econometrics that occurred after the start of our forecast evaluation exercise.

make the lessons of this exercise applicable in practice, we have the further objective of estimating the models and constructing forecasts in a way that is broadly consistent with their use in the Bank of England.

3.1. Estimation data

Our dataset is constructed from a set of real-time forecasts and associated datasets that have been being stored by the Bank of England since 1997. Real-time data for most of the variables required for estimating the two models can be taken from this database directly. However, in some cases there were substantial definitional changes and/or the required series had not been stored in the database over some of the forecast evaluation sample. In such cases, we reconstructed the series manually in a way that was consistent with our objective of using only information that was available prior to each forecast origin.

Our evaluation exercise covers the period 2000Q1–2013Q1. As was discussed in Section 2, we re-estimate COMPASS over this period using a recursive scheme with an estimation sample that begins in 1993Q1. The Statistical Suite is re-estimated at each forecast origin using a rolling window of seven years. This reflects the way in which each model is used in the Bank of England. Under the assumption that all forecasts are 'primitives' (including the model-free *Inflation Report* forecast), this difference in the estimation scheme does not invalidate the statistical methods that we use to evaluate the forecast accuracy.

The *Inflation Report* enjoys a slight informational advantage over the alternative models, due to differences in the timeliness of data. Whereas the *Inflation Report* uses data up to a few days before its publication, both COMPASS and the Statistical Suite are estimated using datasets that are constructed approximately one month ahead of each *Report*. Fig. 2 illustrates the timing of data releases in which earlier, 'benchmark', data are compiled in preparation for the MPC's first meeting during a forecast round, with subsequent data releases being incorporated into the final published forecast. In practice, alternative forecasts produced using COMPASS or the Statistical Suite are based on the early data, so we maintain this convention in our evaluation in order to mimic the real-life conditions in which the models are used.

One issue is that at the benchmark stage, some data for the previous quarter are available (e.g. financial market prices like the exchange rate), but others are not (such as GDP). We fill in this "ragged edge" using the staff's real-time estimates for the missing variables (which are released for the first time between the benchmark and the *Inflation Report*). This ragged edge is treated slightly differently in the estimation of COMPASS and the Statistical Suite, as the data used in the real-time estimation of COMPASS exclude this incomplete quarter, but the data used in the real-time estimation of the Statistical Suite do not. For example, to produce model forecasts that correspond to the November 2012 *Inflation Report*, we use real-time data between 1993Q1 and 2012Q2 to estimate COMPASS and real-time data between 2005Q4 and 2012Q3 to estimate the Statistical Suite (reflecting the seven-year rolling window). We ensure that the information sets that are used to produce the forecasts are the same by imposing the ragged edge together with the estimates of the missing data as judgment in the first quarter of forecasts constructed with COMPASS (using all of COMPASS's shocks in a "full inversion").

3.2. Forecast evaluation data

We evaluate the GDP growth forecasts of COMPASS, the *Inflation Report* and the Statistical Suite against the data in the first Quarterly National Accounts, which are published with a lag of about three months.[13]

One complication when evaluating the inflation forecasts is that the MPC's target measure of inflation changed in December 2003 from the RPIX to the CPI series. Thus, we use RPIX inflation to evaluate the *Inflation Report* forecasts up to 2003Q4, and CPI inflation for forecasts after that date. Consistent with the data used for estimation, we use CPI inflation to evaluate the inflation forecasts from the Statistical Suite and COMPASS.[14]

4. Statistical methods to evaluate forecast accuracy

Before describing the different methods that we use for evaluating the forecast accuracy, we have to introduce some notation. We split the sample, which is of length $T+h$, into an in-sample period of length R and an out-of-sample period of length P, where P is such that $R+P+h=T+h$ and h is the forecast horizon. The outturn of a variable to be predicted is denoted by y_t and the h-period-ahead forecast at origin t is $y_{t+h|t}$. Forecast errors are defined as the difference between outturns and forecasts, $\widehat{u}_{t+h} \equiv y_{t+h} - y_{t+h|t}$.[15] Because both the estimation and the evaluation data are in real time, both forecasts and outturns also depend on the forecast origin, which is suppressed for notational convenience.

4.1. Accuracy of point forecasts

The root mean squared forecast error (RMSFE) is a popular measure of the accuracy of point forecasts. Under the assumption that the loss function is quadratic, the RMSFE at horizon h is given by:

$$\text{RMSFE}^h = \sqrt{\frac{1}{P}\sum_{t=R}^{T}\widehat{u}_{t+h}^2}. \quad (1)$$

13 Our results are robust to using the data vintage that was available in August 2013.

14 Our analysis spans a period over which several changes to VAT occurred. COMPASS is estimated on a measure of inflation that excludes the effects of these changes; see Burgess et al. (2013, Section 8.2) for more details on the estimated effects of VAT changes. Thus, we compare COMPASS inflation forecasts to this series for inflation that excludes the contribution of VAT.

15 Forecast errors also depend on the estimated parameters. That is, the estimation error is captured under the null hypothesis of alternative tests, which means that we adopt the asymptotic framework of Giacomini and White (2006) for conducting inference. That framework is applicable if the parameters are estimated using a small rolling window of observations. However, in a simulation study, Clark and McCracken (2013) document that tests within the Giacomini-White framework perform well for both rolling and recursive schemes.

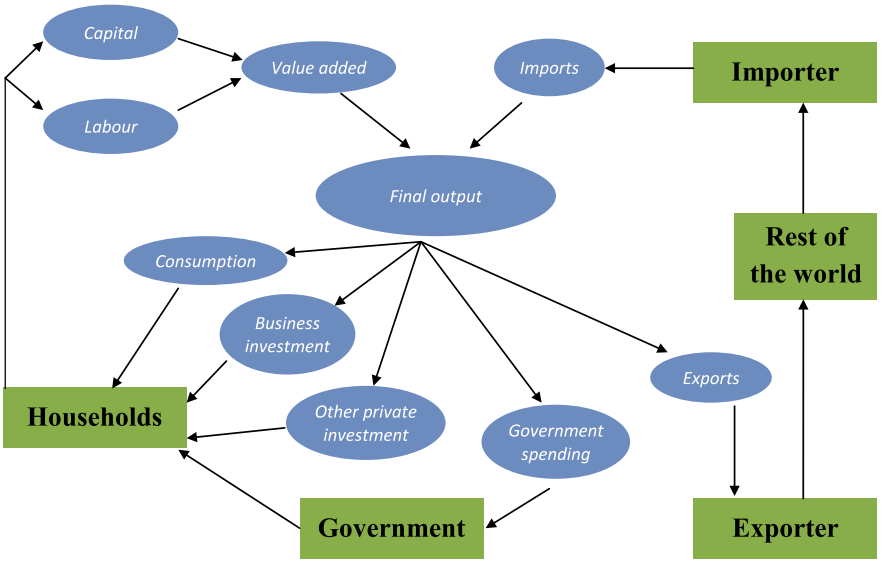


Fig. 1. Overview of COMPASS. The *Central Organizing Model for Projection Analysis and Scenario Simulation* (COMPASS) is the main organizing framework for the *Inflation Report* forecasts at the Bank of England.

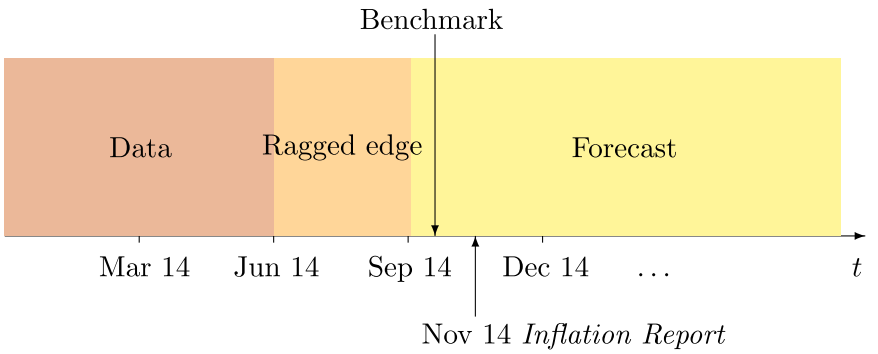


Fig. 2. Timeline of the forecast process at the Bank of England in the case of the November 2014 forecast round.

Diebold and Mariano (1995) proposed the following test statistic for testing whether the RMSFEs from alternative forecasts are significantly different:

$$\mathrm{DM}_h = \frac{1}{\sqrt{P}} \sum_{t=R}^{T} \frac{\widehat{u}^2_{1,t+h} - \widehat{u}^2_{2,t+h}}{\sqrt{\widehat{\Sigma}}}, \tag{2}$$

where $\widehat{\Sigma}$ is an estimate of the long-run variance. Throughout, we use the Newey–West estimator, where the bandwidth is chosen optimally.[16,17] Under the null hypothesis of equal forecast accuracy, DM_h converges to a normal distribution, provided that the loss difference is covariance stationary and has a constant mean and variance.[18]

The Diebold-Mariano test is suitable for assessing the *unconditional* relative predictive abilities of two alternative forecasting models. However, it can also be of interest to investigate whether, for example, one model predicts more accurately in times of high uncertainty while another performs better in normal times. With this objective, Giacomini and White (2006) propose the estimation of the following regression model:

$$\widehat{u}^2_{1,t+h} - \widehat{u}^2_{2,t+h} = \alpha + \beta X_t + e_{t+h}, \tag{3}$$

where X_t contains information that is known at the forecast origin t, such as indicators of economic activity or measures of the global uncertainty. If X_t contains only a constant, the Giacomini and White (2006) test is equal to the test

[16] The slow rate of convergence of heteroskedasticity and autocorrelation consistent covariance estimators means that they may perform poorly in small samples (Haan & Levin, 1996). Fawcett, Koerber, Masolo, and Waldron (2015) addressed this concern by documenting the small sample properties of the Diebold-Mariano test. We find that the Diebold-Mariano test has the correct size and acceptable power in sample sizes like ours.

[17] Clark and McCracken (2014) show that imposing conditioning paths on model-based forecasts induces high-order serial correlation in the forecast errors. We therefore always use an optimally chosen bandwidth to estimate the long-run variance, rather than truncating the bandwidth at $h - 1$, as is often recommended in the forecast evaluation literature Diebold and Mariano (1995).

[18] Non-stationary factors that are common to both forecast errors, such as the global financial crisis, vanish from the loss difference.

of Diebold and Mariano (1995). Under the null hypothesis, two alternative point forecasts are equally accurate *conditional* on X_t.

The test statistic of the *conditional* relative predictive ability test takes the form:

$$\mathrm{GW}_h = P\overline{Z}'\widehat{\Sigma}^{-1}\overline{Z}, \tag{4}$$

where $\overline{Z}$ denotes the vector:

$$\left(\frac{1}{P}\sum_{t=R}^{T}(\widehat{u}^2_{1,t+h}-\widehat{u}^2_{2,t+h}),\ \frac{1}{P}\sum_{t=R}^{T}(\widehat{u}^2_{1,t+h}-\widehat{u}^2_{2,t+h})X_t\right)'. \tag{5}$$

Asymptotically, the Giacomini-White test has a $\chi^2(2)$ distribution.

If H_0 is rejected, it is possible to predict which method will have a lower h-step-ahead loss using current information. In that case, Giacomini and White (2006) propose a simple decision rule: use the first model if the predicted loss is negative and the second model otherwise. More formally, they propose that the first model be used to forecast h steps ahead at time T if $\widehat{\alpha}+\widehat{\beta}X_T < 0$. However, this decision rule assumes that β is constant over time, which is probably not realistic in unstable environments.

Recently, a large body of literature on statistical methods for evaluating the forecast accuracy has emerged. Important recent contributions include the development of statistical tests for the accuracy of density forecasts (e.g. Corradi & Swanson, 2006; Amisano & Giacomini, 2007) and forecast evaluation under instabilities (e.g. Giacomini & Rossi, 2010). We now describe these in more detail.

4.2. Accuracy of density forecasts

While it is straightforward to compare a point forecast to the actual outcome, evaluating a complete density forecast is more difficult. This is because we never observe the entire distribution, but observe only one realization from it. A popular way of overcoming this constraint is to calculate how likely it would be to observe the realized value under the forecast density. This statistic is known as the Probability Integral Transform (PIT). Formally, for a random variable Y_{t+h} and a forecast density constructed at time t, the PIT for the realization y_{t+h} is:

$$z_{h,t} = \int_{-\infty}^{y_{t+h}} f_t(Y_{t+h})dY_{t+h}, \tag{6}$$

so the PIT is the probability of observing a value that is equal to or less than the realised value under the forecast distribution.

If a set of forecast densities offers a good approximation to the true underlying density, then the PITs should be distributed evenly over all percentiles. The intuition is that we would expect an outturn to occur as frequently in practice as the forecast predicted in theory. For example, in a "well calibrated" forecast density — one which matches the underlying distribution correctly — we would expect to see PITs between 0 and 0.2 in one-fifth of outturns. Put differently, the marginal distribution of the PITs is uniform.

A related method of evaluating density forecasts is scoring rules. A scoring rule is a loss function that takes the density forecast and the actual outcome as its arguments. We use the logarithmic scoring rule $\log f(y)$, where f is the density forecast and y is the observed value of the variable in question. The logarithmic score takes a high value if the forecast density assigns a high probability to the actual outturn.

We measure the accuracy of density forecasts over the out-of-sample period using the average logarithmic score, which is defined as:

$$\frac{1}{P}\sum_{t=R}^{T}\log f_t(y_{t+h}). \tag{7}$$

We test whether the logarithmic scores from competing models are significantly different by applying the (unweighted) likelihood ratio test of Amisano and Giacomini (2007). Under the null hypothesis, the two density forecasts $f_{1,t}(\cdot)$ and $f_{2,t}(\cdot)$ perform equally well. The test statistic is given by:

$$\mathrm{AG}_h = \frac{1}{\sqrt{P}}\sum_{t=R}^{T}\frac{\log f_{1,t}(y_{t+h})-\log f_{2,t}(y_{t+h})}{\sqrt{\widehat{\Sigma}}}. \tag{8}$$

Asymptotically, AG_h is distributed as a $N(0,1)$ variable.

4.3. Forecast evaluation in the presence of instabilities

The Diebold-Mariano test assesses relative forecast performances on average over the out-of-sample period. However, relative performances can change over time, as certain events, such as the financial crisis, may affect the predictive content of the variables being forecast. The fluctuation test of Giacomini and Rossi (2010) tests for predictive ability when the predictive content is unstable over time, but changes in a smooth way. Technically, this is achieved by using a kernel. For a rectangular kernel, the fluctuation test amounts to calculating the Diebold-Mariano test over a rolling window of size m, where m is a user-defined bandwidth.

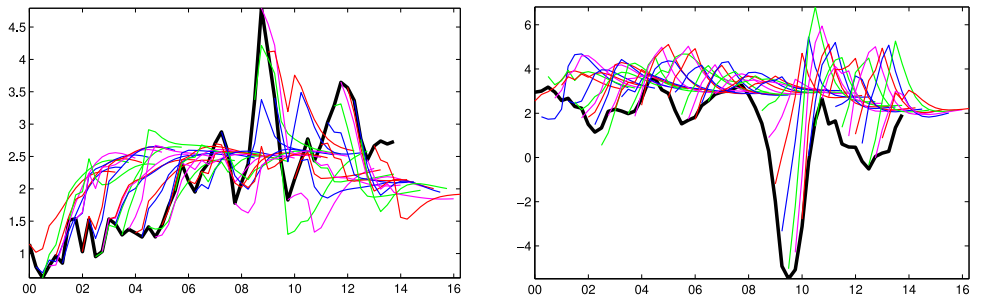
Giacomini and Rossi (2010) test the null hypothesis that both models have equal predictive abilities at each point in time by proposing the fluctuation test statistic:

$$F_{P,h} = \max_t |F_{t,h}|, \tag{9}$$

where:

$$F_{t,h} = \frac{1}{\sqrt{m}}\sum_{j=t-m/2}^{t+m/2-1}\frac{u^2_{1t+h}-u^2_{2t+h}}{\sqrt{\widehat{\Sigma}}},$$
$$t = R+\frac{m}{2},\ldots,T-\frac{m}{2}+1. \tag{10}$$

Under the null hypothesis, the fluctuation test statistic converges to functionals of the Brownian motion. Critical values can be obtained by simulation and are reported in Table 1 of Giacomini and Rossi (2010). In addition to $F_{P,h}$, the time series $F_{t,h}$ itself can be investigated: if it crosses the upper bound, the second model has a superior forecasting performance, while the first model is preferred if the lower bound is crossed.



(a) Inflation (b) GDP growth

Fig. 3. COMPASS point forecasts. *Notes:* Annual inflation and GDP growth ("first final" equivalent) in black, with point forecasts in color: the November rounds are in red, the February rounds in blue, the May rounds in green and the July rounds in magenta. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Forecast performance of COMPASS compared to statistical and judgmental benchmarks

5.1. Descriptive results

We start by graphically comparing the point forecasts from COMPASS, the Statistical Suite, and the *Inflation Report* against the actual outcomes (Figs. 3–5). The inflation projections tend to be attracted towards their underlying mean over each forecast horizon. For the Statistical Suite, this mean is simply the average rate of inflation over the sample period of seven years prior to the forecast origin. This explains why the suite inflation forecasts for one to two years ahead became progressively higher: the actual inflation rate tended to rise over the evaluation window. In the case of COMPASS, the path of inflation over the forecast horizon is driven partly by a natural mean-reversion back to the inflation target of 2%, as is evident in Fig. 3. However, in this larger model, the deviation of inflation from the target is driven by factors that also drive other variables, such as GDP, away from their steady state growth rates. This partly explains why inflation overshoots the target in the pre-crisis period: the forecasts for GDP growth over the same period also overshoot the trend. There is no automatic mean-reversion of the *Inflation Report* forecasts as the projections reflected the MPC's judgment about the evolution of the state of the economy, at the time of the forecast. The Committee's forecasts of inflation proved to be quite accurate before the crisis, while the post-crisis period saw frequent underprediction.

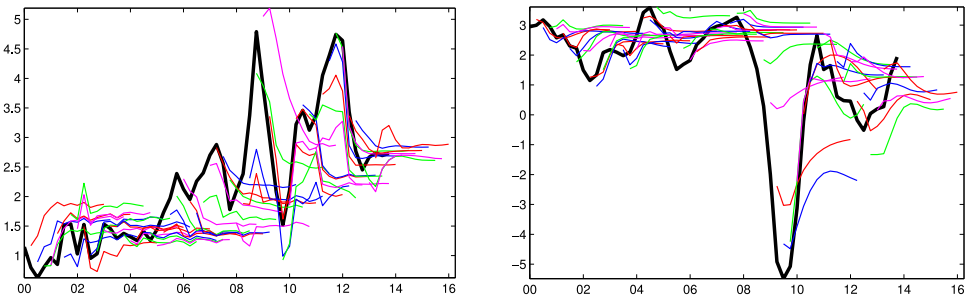
All GDP growth forecasts overshoot the outturns at some point in the evaluation period. This is particularly marked for COMPASS in Fig. 3; its tendency to over-predict growth is accentuated after the crisis, relative to both its earlier performance and those of the other forecast methods. To understand why this happens, Section 6 considers additional information and judgment that is incorporated into the *Inflation Report* projections, but is not necessarily incorporated into forecast models.

In COMPASS and the Statistical Suite, the speed with which the GDP is forecast to recover is affected by each model's estimate of the long-run trend rate of growth. In purely statistical models with a trend growth rate (including most models in the suite), growth will tend to revert towards the long-run equilibrium within a few quarters of the forecast origin. In more structural models, such as COMPASS, the phenomenon of mean-reversion will still hold, but the exact speed of convergence will depend on the nature of the underlying shocks at the forecast origin.[19] Thus, the accuracy of these GDP growth forecasts depends on whether the estimate of the long-run growth is correct. In COMPASS, trend growth is calibrated based on sample data. Against a backdrop of sustained weakness in the aftermath of the crisis, it is not surprising therefore that the model significantly overpredicts growth.[20] Models with trend growth rates that can vary over time are likely to be more robust to the kinds of shocks wrought by the financial crisis. In particular, the Statistical Suite incorporates measures of the trend estimated over the seven-year period prior to the forecast origin. This improves the forecast performance, especially following the crisis, as can be seen in Fig. 4b. The estimated trend growth rate falls with each successive forecast origin after 2008, as the crisis begins to dominate the sample period, which in turn pushes down on the GDP growth forecasts. This result chimes with other papers in the literature that stress the importance of specifying trends correctly for the estimation and empirical fit of DSGE models (e.g. Ferroni, 2011).

There may also be reasons to believe that GDP growth will take a longer — or shorter — time to return to its trend

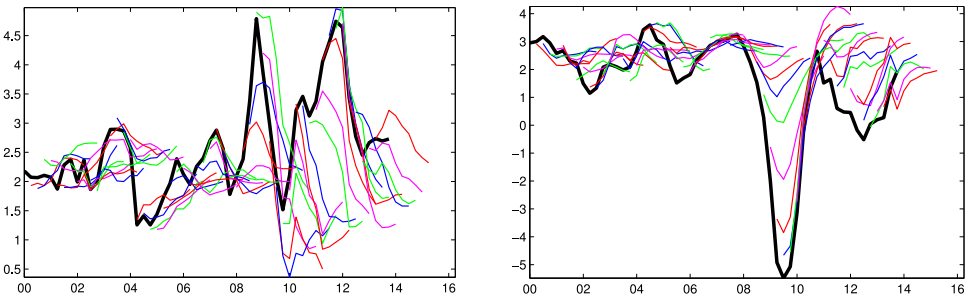
19 Like all models of its class, there are two types of mean reversion in COMPASS. First, all variables have constant long-run growth rates. The long-run growth rate of GDP is set equal to the sample average in each recursively estimated variant. These sample averages tend to exceed the growth rates observed in the data after the crisis. Secondly, although COMPASS attributes a material part of the fall in GDP over the crisis to a permanent productivity shock, the rest of the fall in GDP is identified as having been driven by temporary shocks (and negative demand shocks, like a domestic risk premium shock, in particular). These shocks unwind over the forecast, so there is some mean reversion built into the level of GDP in COMPASS, as well as the growth rate. Section 5.5 of Fawcett et al. (2015) discusses the shocks identified by COMPASS over the crisis period in a bit more detail.

20 Trend productivity growth in COMPASS is calibrated using data from the twenty-year period up to the financial crisis. However, as Barnett, Batten, Chiu, Franklin, and Sebastiá-Barriel (2014) note, there was a striking decline in growth after the crisis, which could lead to persistent forecast failure in models that are not robust to such shifts. With this in mind, we also explored a measure of the time-variation in trend growth that was presented by Fawcett et al. (2015).



(a) Inflation (b) GDP growth

Fig. 4. Statistical Suite point forecasts. *Notes:* See the notes to Fig. 3.



(a) Inflation (b) GDP growth

Fig. 5. *Inflation Report* point forecasts. *Notes:* See the notes to Fig. 3.

rate than would normally be the case. This alludes to the role of expert judgment, and additional information that is not normally incorporated into a forecast model, in forecasting. The *Inflation Report* forecasts in Fig. 5b illustrate the former. In the immediate aftermath of the crisis, the MPC predicted a rapid recovery in GDP growth rates, which failed to materialize. As Hackworth, Radia, and Roberts (2013) noted, this forecast failure broadly reflected three judgments of the Committee that turned out to be incorrect: world growth was unexpectedly weak; credit supply was unexpectedly tight and uncertainty elevated; and import and energy costs were unexpectedly high. These factors drove a wedge between the MPC's anticipated return to trend growth and the actual path.

Thus, the broad point from this and the *Inflation Report* forecasts is that a deeper understanding of the factors that push GDP growth away from its equilibrium can sometimes be a help when forecasting. This is likely to be particularly important following a large shock like the financial crisis. Motivated by these observations, Section 6 assesses how augmenting the set of model observables with survey measures of expected GDP growth can be effective at improving the forecast accuracy when the deviations from trend are particularly large and persistent.

5.2. Point forecast evaluation

We evaluate the point forecast accuracy of each of the alternative forecast methods using the statistical methods described in Section 4.1. We compute the root mean squared forecast errors (RMSFEs) for GDP growth and inflation at different horizons and test for significant differences in forecast accuracy between forecast methods, then investigate whether these differences were predictable given the information available at the time when each forecast was made. Here, we concentrate on performances across the sample as a whole; Section 5.4 then assesses whether the forecast performances change over the sample.

The RMSFEs for all forecast methods are shown in Figs. 6a and 6b for inflation and GDP growth respectively. Looking first at the near-term forecasts (up to three quarters ahead), the *Inflation Report* forecasts are more accurate than any of the model-based forecasts. Further ahead, though, the rankings reverse, in particular for inflation, where the RMSFEs of the *Inflation Report* forecasts are highest.

For GDP growth, these differences are statistically significant at a horizon of one year (Table 1), with the *Inflation Report* and Statistical Suite GDP forecasts outperforming COMPASS significantly. In contrast, the differences in RMSFEs for inflation are not statistically significant (Table 1).

Statistical significance notwithstanding, the finding that judgmental forecasts tend to perform better at short horizons while model-based forecasts perform better at longer horizons is consistent with many papers in the literature. Del Negro and Schorfheide (2012) find that the forecasts for inflation and GDP growth from the Smets-Wouters model are more accurate than the Federal Reserve Board staff's judgmental Greenbook forecasts at horizons of three quarters or longer, and Edge et al. (2010) find the

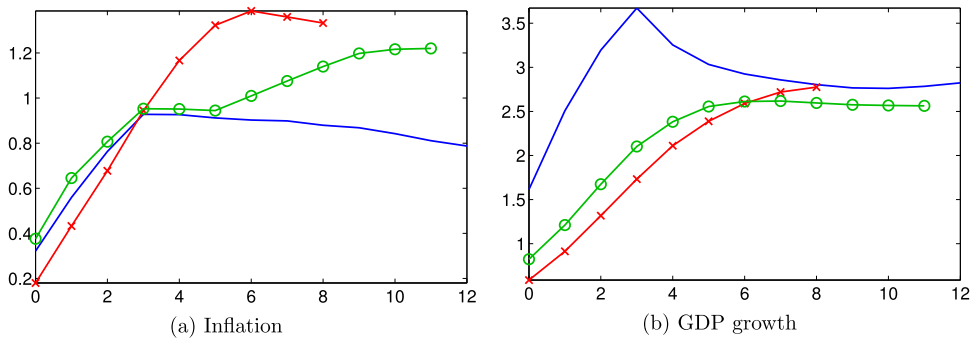


Fig. 6. Root mean squared forecast errors. *Notes:* The COMPASS forecasts are in blue, the Statistical Suite forecasts in green with circles and the *Inflation Report* forecasts in red with crosses.

Table 1
Diebold–Mariano test of equal relative forecasting abilities at a horizon of one year.

	COMPASS	IR	Stat. Suite
(a) Inflation			
COMPASS	.	−0.806	0.135
IR	.	.	0.939
Stat. Suite	.	.	.
(b) GDP growth			
COMPASS	.	1.815*	2.128**
IR	.	.	−0.792
Stat. Suite	.	.	.

Notes: The table shows test statistics. A Newey–West estimator with an optimally-chosen bandwidth is used to estimate the long-run variance. If the optimal bandwidth exceeds $2(h-1)$, $2(h-1)$ is used instead. ** (*) indicates statistical significance at the 5% (10%) level. A value that is smaller(larger) than zero indicates that the model in the corresponding row generates more (less) accurate forecasts than that listed in the corresponding column.

Table 2
Giacomini-White test of equal conditional relative forecasting abilities at a horizon of one year (oil price).

	COMPASS	IR	Stat. Suite
(a) Inflation			
COMPASS	.	1.273	2.813
IR	.	.	2.085
Stat. Suite	.	.	.
(b) GDP growth			
COMPASS	.	7.019^{++}	17.379^{++}
IR	.	.	1.227
Stat. Suite	.	.	.

Notes: The table shows test statistics. The test functions are $(1, X_t)$, where X_t is the quarterly growth rate of the oil price. A Newey–West estimator with an optimally chosen bandwidth is used to estimate the long-run variance. A plus (minus) sign indicates that the test rejects equal predictive ability and that the method in the row has a larger (smaller) predicted loss on average. + (++) (− (−−)) indicates significance at the 10% (5%) level.

same result using the Federal Reserve Board model. Petrova (2013) finds that the Smets and Wouters (2003) DSGE model applied to UK data outperforms the *Inflation Report* for forecasting inflation at some horizons.

The Diebold-Mariano test assesses the *unconditional* relative predictive ability of two alternative forecasting models. however, it is also of interest to investigate

Table 3
Giacomini-White test of equal conditional relative forecasting abilities at a horizon of one year (corporate credit spread).

	COMPASS	IR	Stat. Suite
(a) Inflation			
COMPASS	.	1.629	0.027
IR	.	.	2.043
Stat. Suite	.	.	.
(b) GDP growth			
COMPASS	.	3.733	7.838^{++}
IR	.	.	0.933
Stat. Suite	.	.	.

Notes: See the notes to Table 2, except that X_t is the corporate credit spread.

whether the relative performances vary *conditional* on information that was available at the forecast origin. Motivated by the substantial tightening in credit conditions over period of the financial crisis and the volatility of oil prices between 2007 and 2009, we test whether the investment-grade UK corporate bond spread and (Brent crude) oil price growth can predict differences in forecasting accuracy among our methods at the one-year horizon. Tables 2 and 3 report Giacomini and White (2006) test statistics.[21] We find that both variables can predict loss differences for GDP growth one year ahead in several of the comparisons.[22] This suggests that, conditional on the credit spread, the Statistical Suite is more accurate than COMPASS, which does not include a measure of financial conditions.

5.3. Density forecast evaluation

Probability Integral Transforms (PITs) are a useful method for assessing whether or not a density forecast is consistent with the observed frequency of outturns. Figs. 7

[21] Theoretically, a rejection of the (unconditional) Diebold-Mariano test should imply a rejection of the conditional test. However, this is not always what we observe. As was discussed by Giacomini and White (2006), the most likely explanation for this is that there are size distortions of the Diebold-Mariano test and a sensitivity to the choice of the lag length.

[22] Although these are not reported here, house prices, the VIX and the sterling ERI can also predict forecast accuracy differences in GDP growth.

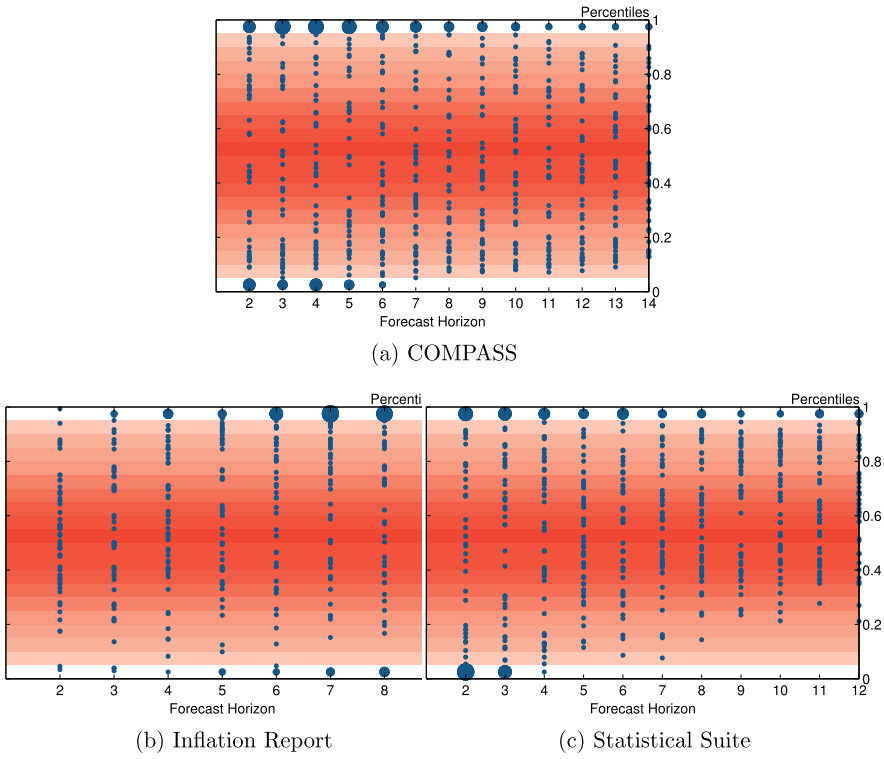


Fig. 7. Probability integral transforms for inflation.

and 8 illustrate the PITs for inflation and GDP growth, respectively, for our forecasting methods. Each dot shows the probability — under the forecast distribution — of observing an outturn that is equal to or lower than the inflation or GDP growth rate that came to pass. The size of the dots is proportional to the frequency of observing a given probability score. If the density forecasts are accurate, we would expect the dots to be spaced roughly evenly between zero and one.

At shorter horizons, COMPASS tends to underpredict inflation (Fig. 7), as is illustrated by the concentration of observations in the upper tail of the distribution. Furthermore, at short forecast horizons, too many PITs are located in either the highest or lowest percentile buckets, which indicates that COMPASS understates the uncertainty around the inflation forecast, and the same appears to be true of the Statistical Suite. However, the one-quarter-ahead inflation forecasts in the *Inflation Report* tend to show the opposite problem, as they overstate the uncertainty around the point forecast. At longer horizons, the inflation PITs for COMPASS and the Statistical Suite seem to be spaced more evenly between zero and one.

Fig. 8 illustrates that all of our forecasting methods overestimate GDP growth, as there are very few outturns in the upper percentile buckets. Overall, at longer horizons, the GDP density forecasts from the Statistical Suite appear to be calibrated the best, with a more even spread of PITs between zero and one.

Scoring rules are a convenient way of summarizing the information that is contained in a density forecast in a single number. A higher score indicates that the forecast density tends to assign a high probability to the realized outturns (Section 4.2). Fig. 9 reports logarithmic scores at different forecast horizons for the densities from each of our forecasting methods. We find that the *Inflation Report* has relatively more accurate density forecasts for inflation and GDP growth at shorter horizons, and the most accurate densities one quarter ahead. At a forecast horizon of one year, the inflation density forecasts from the Statistical Suite are significantly more precise than those of the other methods (Table 4).

For GDP growth, the *Inflation Report* and the Statistical Suite are more accurate than COMPASS at near-term horizons. At longer horizons, the Statistical Suite forecasts are substantially worse than other methods. This is due largely to the financial crisis period, during which some of the GDP growth outturns were a long way into the lower tail of the Suite's probability forecast. Given the nature of the logarithmic transformation of these small probabilities, the penalty applied to the Suite over this period is material.

5.4. Instabilities in forecast performance

Above, we evaluated the forecasting performances of our different methods across the entire evaluation sample. We now turn to an investigation of the way in which that performance has varied over time, paying particular attention to the financial crisis as a cause of instability.

Fig. 10 reports average logarithmic scores for the pre- and post-crisis periods separately. For both inflation and GDP growth, log scores (looking across all three forecast methods) are generally lower in the earlier period than

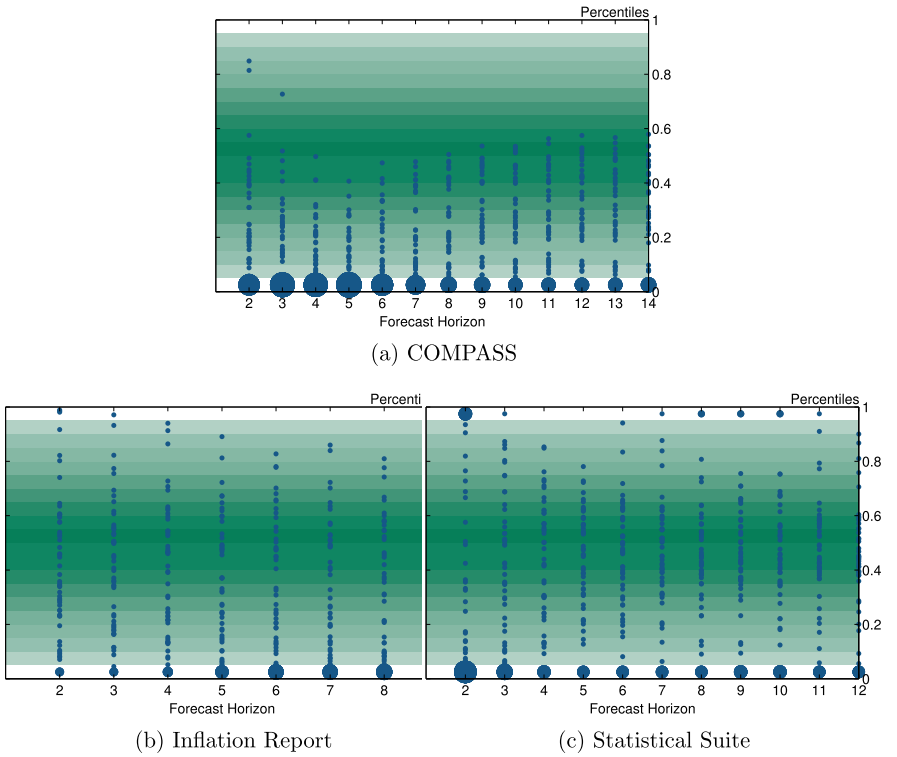


Fig. 8. Probability integral transforms for GDP growth.

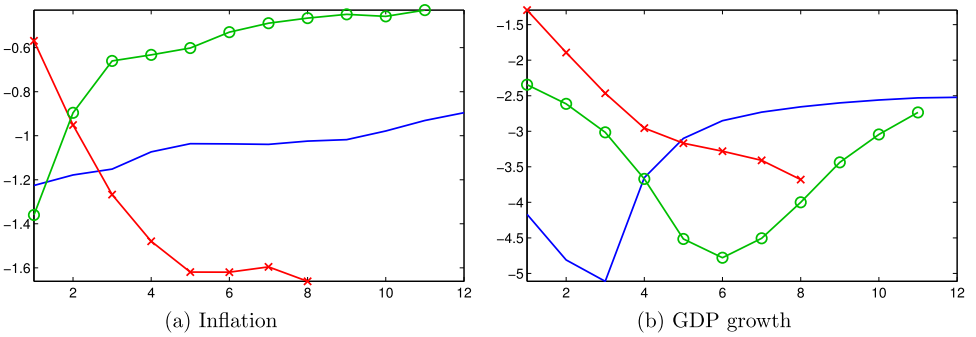


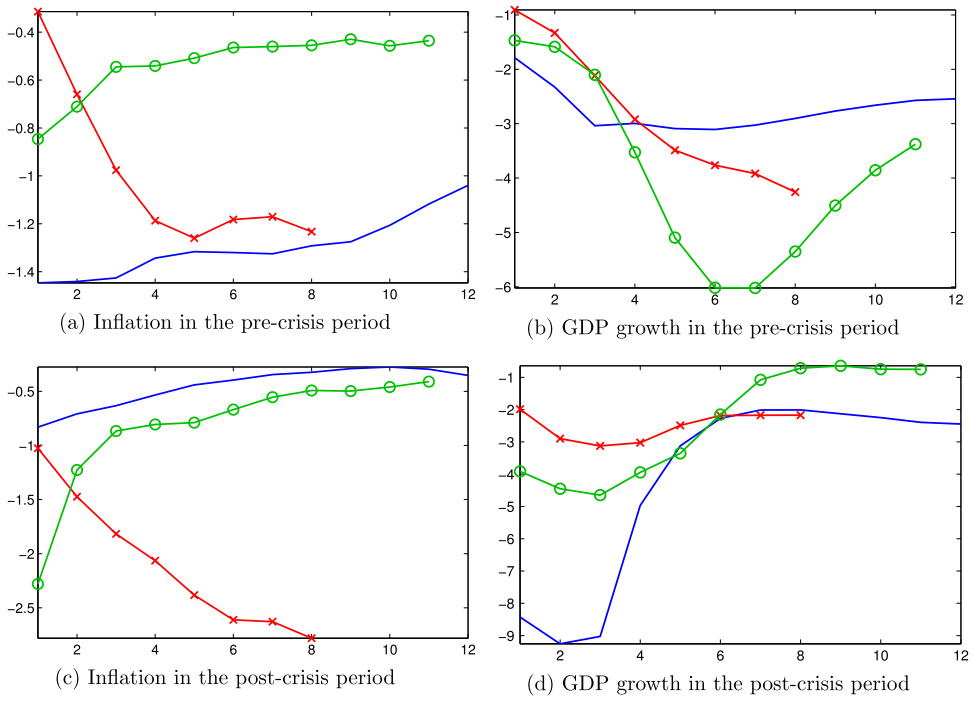
Fig. 9. Average logarithmic scores. *Notes:* The COMPASS forecasts are in blue, the Statistical Suite forecasts in green with circles and the *Inflation Report* forecasts in red with crosses.

the later, which indicates a general deterioration in density forecast performance after the crisis. At short forecast horizons, the difference in log scores between GDP growth density forecasts that implicitly or explicitly take into account a broad information set (the Statistical Suite and the *Inflation Report*) and those which do not (COMPASS) is larger in the post-crisis period. This observation suggests that conditioning on a broad and timely set of indicators can improve the forecast accuracy, especially in times of heightened volatility, a theme that we explore in more detail below.[23]

It is also informative to investigate how the logarithmic scores have changed over time. Fig. 11 calculates the logarithmic scores for GDP growth and inflation density forecasts over a rolling window of four years at horizons of one and two years. For GDP growth at both horizons, the logarithmic score falls sharply at the onset of the financial crisis, reflecting a substantial loss in accuracy of the density forecasts. The pattern for the inflation density forecasts is much less clear, although the scores of the inflation density forecasts for the *Inflation Report* did decrease sharply during the crisis, while those for the other methods under consideration did not.

Fig. 12 reports test statistics from the fluctuation test: the difference between the average MSFE calculated over a rolling window of 28 quarters standardized by an estimate

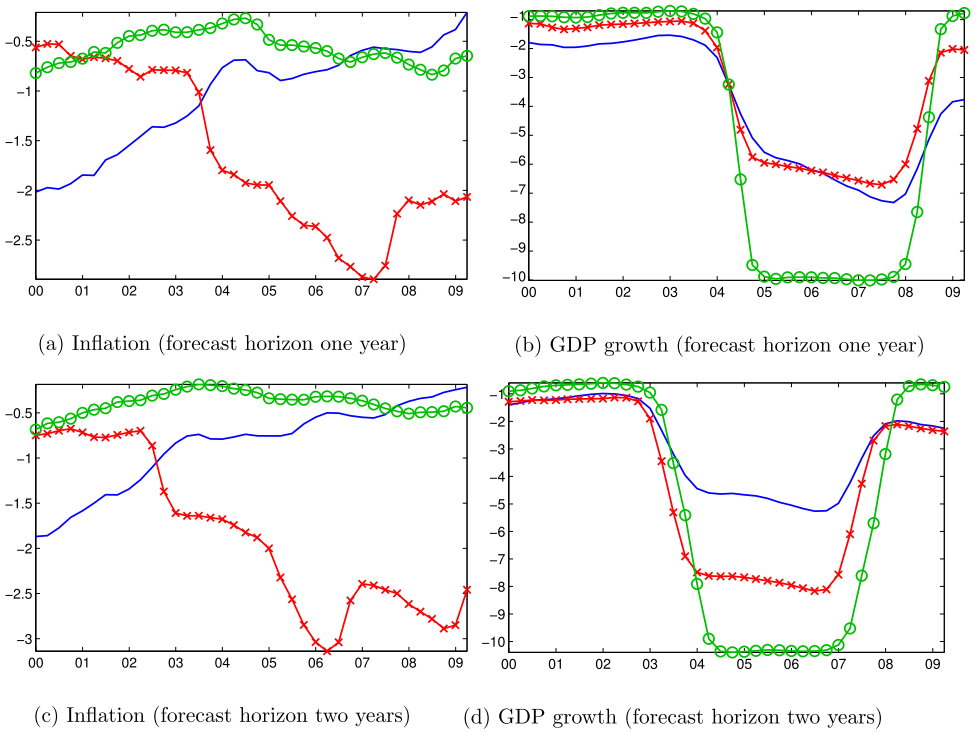
[23] The seven-year rolling widow estimation for the Statistical Suite is also important for delivering a relatively good post-crisis forecast performance.



(a) Inflation in the pre-crisis period (b) GDP growth in the pre-crisis period

(c) Inflation in the post-crisis period (d) GDP growth in the post-crisis period

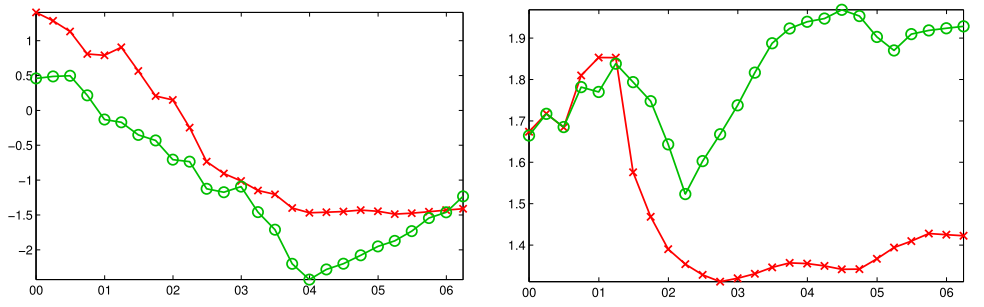
Fig. 10. Average logarithmic scores for the pre- and post-crisis periods. *Notes:* The COMPASS forecasts are in blue, the Statistical Suite forecasts in green with circles and the *Inflation Report* forecasts in red with crosses.



(a) Inflation (forecast horizon one year) (b) GDP growth (forecast horizon one year)

(c) Inflation (forecast horizon two years) (d) GDP growth (forecast horizon two years)

Fig. 11. Average logarithmic scores calculated over a rolling window of size four years. *Notes:* The x-axis indicates the start of the window. The COMPASS forecasts are in blue, the Statistical Suite forecasts in green with circles and the *Inflation Report* forecasts in red with crosses.



(a) Inflation (b) GDP growth

Fig. 12. Fluctuation test statistics for a forecast horizon of one year. *Notes:* The size of the rolling window is seven years. The x-axis indicates the start of the window. All models are evaluated relative to COMPASS. A positive (negative) value of the test statistic means that COMPASS performs worse (better). The (relative) Statistical Suite forecasts are shown in green with circles and the *Inflation Report* forecasts are in red with crosses. Critical values for a two-sided test are 2.779 (2.5) at the 5% (10%) level (Table 1 of Giacomini and Rossi, 2010).

Table 4
Amisano-Giacomini test of equal density forecasts at a horizon of one year.

	COMPASS	IR	Stat. Suite
(a) Inflation			
COMPASS	.	0.680	−1.761*
IR	.	.	−1.806*
Stat. Suite	.	.	.
(b) GDP growth			
COMPASS	.	−1.556	0.055
IR	.	.	0.642
Stat. Suite	.	.	.

Notes: The table shows test statistics. A Newey–West estimator with an optimally chosen bandwidth is used to estimate the long-run variance. If the optimal bandwidth exceeds $2(h-1)$, $2(h-1)$ is used instead. ** (*) indicates statistical significance at the 5% (10%) level. A value that is smaller (larger) than zero indicates that the model in the corresponding column generates more (less) accurate forecasts than that listed in the corresponding row.

of the long-run variance. The test statistics reported are relative to COMPASS, such that a negative value of the test statistic means that this model performs better in terms of average MSFEs. A casual inspection of Fig. 12 indicates that the relative forecasting performance has varied over the sample period. For inflation, the relative performances of the Statistical Suite and the *Inflation Report* tended to decline over the sample relative that of COMPASS. Indeed, the test statistic suggests that COMPASS outperformed the other two methods under consideration for a large part of our sample. The picture is different when it comes to GDP growth. COMPASS is generally outperformed, and the performance of the Suite improves over time in relative terms. While none of these fluctuation test statistics are statistically significant, this is due at least in part to the relatively small sample size.

6. The role of off-model information

This section investigates whether the forecast performance of COMPASS can be improved when it is augmented with different types of judgment and off-model information that are explained in detail in what follows.

6.1. Types of judgment and off-model information

Some of the judgments that we apply to COMPASS closely follow the way in which DSGE models are used at policy institutions, where the forecasts are often conditional on the future paths of particular variables such as the policy rate. The individual sets of off-model information that we consider are as follows:

1. **Survey growth expectations:** The survey-augmented version (COMPASS$^{GDP^e}$) adds a survey measure of one-year-ahead growth expectations to the set of observables that are used to identify the state in the baseline model.[24] Because COMPASS contains more shocks than observables, this does not require new shocks to be added to the model. COMPASS$^{GDP^e}$ uses the parameter estimates from COMPASS. Note that, as was discussed in Section 2, we estimate COMPASS using information up to and including the quarter before the forecast origin. Consistent with that, it is the model's three-quarter-ahead forecast of GDP growth that coincides with the lagged one-year-ahead survey expectations measure.
2. **Inflation nowcasts:** The nowcast-augmented version of COMPASS (COMPASS$^{\pi^n}$) constrains the current and one-quarter-ahead inflation forecast to match a nowcast produced by Bank staff. Inflation nowcasts are applied using a final output markup shock that is broadly consistent with the way in which this information is treated at the Bank of England. Section 6.3 also examines the effect of applying the inflation nowcasts using all of the shocks in the model (rather than only the markup shock).
3. **Conditioning path of the policy rate:** In the COMPASS^R variant, the interest rate is constrained to follow a particular path that is derived from market expectations. Interest rate paths are applied using a

[24] This survey measure is taken from a quarterly survey of external forecasters' one-, two- and three-year-ahead expectations for a small number of macroeconomic variables that appear in each *Inflation Report*.

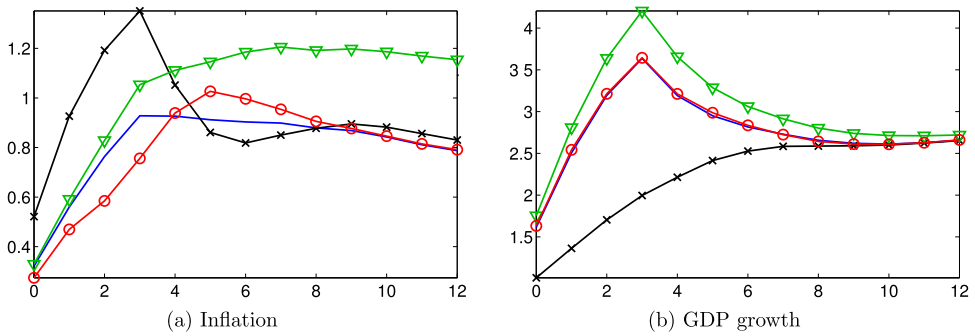


Fig. 13. The role of off-model information: root mean squared forecast errors. *Notes:* The COMPASS forecasts are in blue, the COMPASS$^{GDP^e}$ forecasts in black with crosses, the COMPASS$^{\pi^n}$ forecasts in red with circles, and the COMPASS^{R} forecasts in green with triangles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

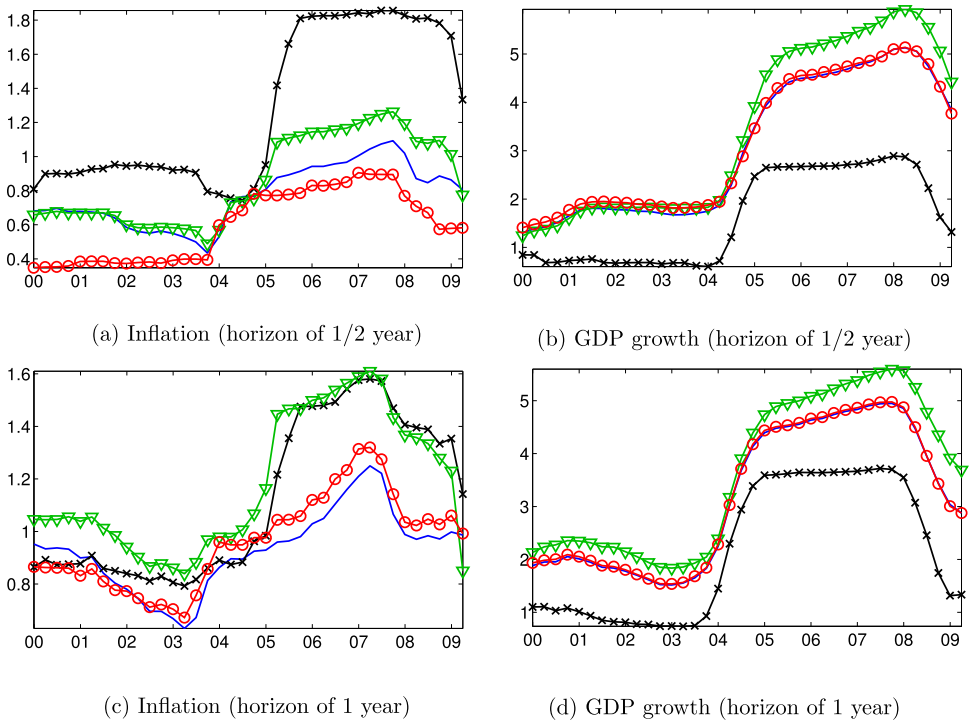


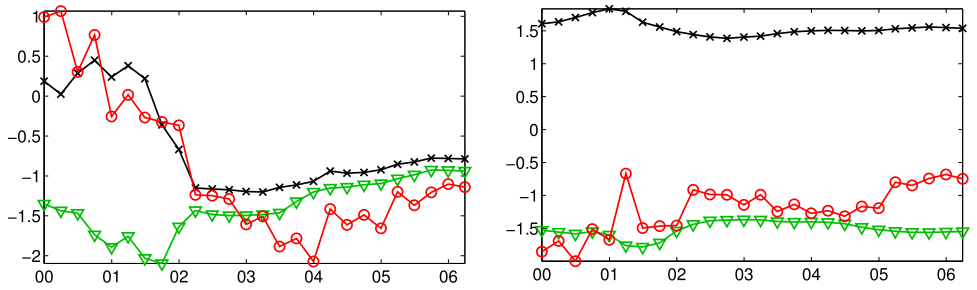
Fig. 14. The role of off-model information: Time-varying root mean squared forecast errors at forecast horizons of 1/2 and one year. *Notes:* The window length is four years. The x-axis shows the start of the window. The COMPASS forecasts are in blue, the COMPASS$^{GDP^e}$ forecasts in black with crosses, the COMPASS$^{\pi^n}$ forecasts in red with circles, and the COMPASS^{R} forecasts in green with triangles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

series of monetary policy shocks. Section 6.3 also explores the application of the policy rate conditioning path using: (a) all of the shocks in the model; and (b) the exponential tilting method of Robertson et al. (2005).

6.2. Off-model information and forecast performance

We evaluate the point forecast accuracy of each of the alternative forecast methods using the statistical methods described in Section 4.1. We compute the RMSFEs for GDP growth and inflation at different horizons over the out-of-sample period, and test for significant differences in forecast accuracy between alternative sets of off-model information.

RMSFEs are shown in Figs. 13 and 14. Imposing inflation nowcasts improves the forecast performance of inflation relative to the model without additional judgments up to one year, but the RMSFE increases a bit further out. The improvement in the short-run forecast performance is particularly evident in the early and later parts of the out-of-sample period (Fig. 14a). This point is also echoed by the fluctuation test statistic reported in Fig. 15a. As above, the differences in MSFE are relative to COMPASS, such



(a) Inflation (b) GDP growth

Fig. 15. The role of off-model information: Fluctuation test statistic for a forecast horizon of one year. *Notes:* The size of the rolling window is seven years. The x-axis indicates the start of the window. All models are relative to COMPASS, where a positive number means that COMPASS performs worse. The COMPASS$^{GDP^e}$ forecasts are in black with crosses, the COMPASS$^{\pi^n}$ forecasts in red with circles, and the COMPASS^{R} forecasts in green with triangles. Critical values for a two-sided test are 2.779 (2.5) at the 5% (10%) level (Table 1 of Giacomini and Rossi, 2010).

Table 5
The role of off-model information: Diebold–Mariano test of equal relative forecasting abilities at a horizon of one year.

	COMPASS	COMPASS$^{GDP^e}$	COMPASS^{R}	COMPASS$^{\pi^n}$
(a) Inflation				
COMPASS	.	−0.711	−1.162	−0.394
COMPASS$^{GDP^e}$	.	.	−0.683	0.714
COMPASS^{R}	.	.	.	1.185
COMPASS$^{\pi^n}$	.	.	.	.
(b) GDP growth				
COMPASS	.	1.921*	−1.801*	−1.341
COMPASS$^{GDP^e}$	.	.	−1.891*	−1.926*
COMPASS^{R}	.	.	.	1.768*
COMPASS$^{\pi^n}$	.	.	.	.

Notes: The table shows test statistics. A Newey–West estimator with an optimally chosen bandwidth is used to estimate the long-run variance. If the optimal bandwidth exceeds $2(h-1)$, $2(h-1)$ is used instead. ** (*) indicates statistical significance at the 5% (10%) level. A value that is smaller (larger) than zero indicates that the model in the corresponding row generates more (less) accurate forecasts than that listed in the corresponding column.

that a positive value of the test statistic indicates that the model under consideration outperforms COMPASS. At the beginning of the sample period in particular, we find that the fluctuation test statistic is positive (but not statistically significant) for COMPASS$^{\pi^n}$, meaning that this variant performs better than the baseline. The accuracy of GDP growth forecast is hardly affected by imposing inflation nowcasts (Fig. 13b).

In contrast, incorporating survey growth expectations produces better GDP growth forecasts than the other versions (Fig. 13b), and the differences are statistically significant at a horizon of one year (Table 5). Similar conclusions can be drawn by inspecting the fluctuation test statistic in Fig. 15b, which takes a positive (but not statistically significant) value for COMPASS$^{GDP^e}$ throughout the sample period. This improvement in forecast accuracy for GDP growth is driven primarily by the crisis period, where the gap in RMSFE between the variant with survey expectations and the other models widens, especially for short horizons (Figs. 14b, 14d).

One reason for this observation is that the additional information contained in the survey forecasts leads COMPASS to appeal to a different mix of underlying shocks for explaining the evolution of the financial crisis, compared to the baseline model. It finds that the data are explained better by more persistent shocks — including a permanent shock to productivity — which depress GDP growth for longer and thus slow down mean-reversion.

At the same time, the pattern of shocks identified by COMPASS$^{GDP^e}$ introduces large forecast errors for inflation at a horizon of around one year. This finding is driven in part by the crisis period, where that model predicted subdued GDP growth correctly but failed to anticipate inflation rates above the target (Figs. 14a, 14c).[25]

[25] Recent work by Galvo, Giraitis, Kapetanios, and Petrova (2016) and Petrova, Kapetanios, Masolo, and Waldron (2017) used time-varying estimation techniques in order to capture structural breaks better. The exercise that we conduct in this paper relies on a recursive estimation approach which allows for a degree of time variation in the parameter estimates. The working paper version of this work (Fawcett et al., 2015) also experiments with a time-varying calibration of trend growth. We find that allowing for time variation in trends does not improve the forecasting accuracy relative to our survey-expectations-augmented version of COMPASS. The differences in inflation forecasting performances are not statistically significant, while COMPASS$^{GDP^e}$ produces significantly more accurate point and density forecasts for one-year-ahead GDP growth (see Tables 1, 2, 5, and 6 of Fawcett et al., 2015).

Table 6
The role of off-model information: Diebold–Mariano test of equal relative forecasting abilities at a horizon of one year when varying the method of applying conditioning information.

	COMPASS	COMPASS^R	COMPASS^R (all)	COMPASS$^{\pi^n}$	COMPASS$^{\pi^n}$ (all)
(a) Inflation					
COMPASS	.	−1.162	−0.814	−0.394	−0.208
COMPASS^R	.	.	0.059	1.185	1.090
COMPASS^R (all)	.	.	.	0.991	1.208
COMPASS$^{\pi^n}$	.	.	.	.	−0.029
COMPASS$^{\pi^n}$ (all)	.	.	.	.	.
(b) GDP growth					
COMPASS	.	−1.801*	2.394**	−1.341	−0.151
COMPASS^R	.	.	2.032**	1.768*	1.584
COMPASS^R (all)	.	.	.	−2.834**	−2.673**
COMPASS$^{\pi^n}$	.	.	.	.	0.098
COMPASS$^{\pi^n}$ (all)	.	.	.	.	.

Notes: The table shows test statistics. A Newey–West estimator with an optimally chosen bandwidth is used to estimate the long-run variance. If the optimal bandwidth exceeds $2(h-1)$, $2(h-1)$ is used instead. ** (*) indicates statistical significance at the 5% (10%) level. A value that is smaller (larger) than zero indicates that the model in the corresponding row generates more (less) accurate forecasts than that listed in the corresponding column. "(all)" indicates that the conditioning information is applied using all shocks.

A general conclusion from evaluating forecasts that have been augmented with off-model nowcasts and survey data is that incorporating relevant and timely off-model information can be useful, especially in periods of heightened volatility. In the same vein, Del Negro and Schorfheide (2012) demonstrate that a version of the Smets and Wouters (2003) model that is modified to incorporate financial frictions forecasts GDP growth during the financial crisis more accurately than the equivalent model without these frictions. It is possible that the survey expectations measure that we use is picking up some of the same information as the measure of credit spreads that is used by Del Negro and Schorfheide (2012) in their model. At the same time, it is interesting to note that we do not change the structure of the model, for instance by adding financial frictions. We only add an extra observable, which we can do without changing the set of shocks because COMPASS has more shocks than observable variables. Thus, the differences in our forecasts reflect the difference in the set of shocks that are filtered in by the model, given the extra information, but the transmission of the shocks remains the same.

Finally, the model that is conditioned on the future paths of interest rates that are expected by market participants produces less accurate inflation and GDP forecasts than the other models, with the exception of inflation at short horizons. This observation holds over both the full out-of-sample period and different episodes (Figs. 13, 14). However, as we will document in the remainder of this section, this result is driven in part by the way in which the conditioning information is applied.

So far, conditioning paths for interest rates (and inflation nowcasts) have been applied using a monetary policy (final output markup) shock. We will now investigate the extent to which this is important.

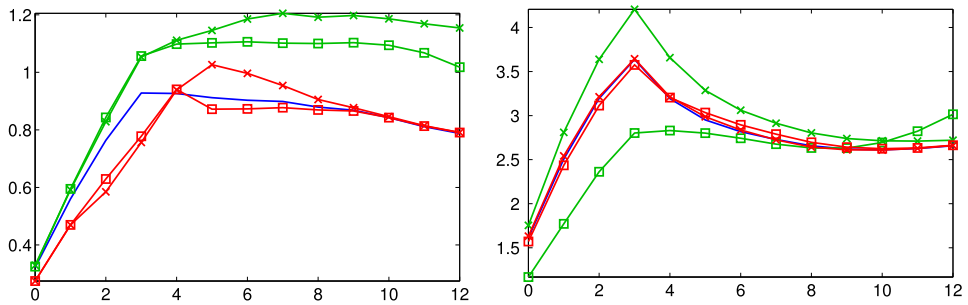
6.3. Alternative ways of imposing off-model information

When we impose conditioning paths for interest rates using a monetary policy shock, we are making two judgemental decisions at once: one concerning the course of the short term rate, and a second one regarding the source of its deviation from an unconditional forecast. This is common practice, and reflects a genuine appreciation of the driving force behind the discrepancy between the unconditional forecast and, say, the market-implied profile for interest rates.

From the perspective of a pure forecasting exercise, though, we can also explore alternative approaches. Fig. 16 reports RMSFEs from an exercise in which we let the Kalman filter select the minimal-shock-variance combination from among the full set of shocks in COMPASS. From an economic perspective, this amounts to relaxing the idea that the difference between the unconditional forecast for the monetary policy rate and the market-implied profile can be ascribed entirely to a different view of how monetary policy will evolve. In contrast, imposing the conditioning information using all shocks allows market participants to have a different view of the outlook of the economy, as well as of monetary policy.

When we apply this approach to inflation nowcasts, we find that both methods of applying the off-model information reduce the RMSFEs for inflation at short horizons relative to the standard version of COMPASS. At horizons of longer than one year, the version in which the nowcast is applied using all shocks is more accurate for forecasting inflation than both the case in which the same information is applied using a final output markup shock and the baseline model. However, these differences are not statistically significant (Table 6). The fact that the shock selection matters less for inflation than for the policy rate can be explained at least in part by the preponderance of markup shocks in explaining the inflation variability in New Keynesian DSGE models (and, therefore, the extent to which this shock is 'selected' in the full inversion).

The results are more interesting when it comes to imposing a profile for the policy rate. In this case, using all shocks improves the forecast performance relative to the variant where the same information is applied using the monetary policy shock. For GDP growth, applying the interest rate path using all shocks also outperforms the baseline model where no additional information is imposed.



(a) Inflation (b) GDP growth

Fig. 16. The role of off-model information: root mean squared forecast errors when varying the method of applying conditioning information. *Notes:* The COMPASS forecasts without any conditioning paths imposed are in blue, the COMPASS$^{\pi^n}$ forecasts using the final output markup shock (all shocks) are in red with crosses (squares) and the COMPASS^{R} forecasts using the monetary policy shock (all shocks) are in green with crosses (squares). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

These differences in predicting GDP growth are statistically significant at a horizon of one year (Table 6). Intuitively, using all shocks captures the idea that economic agents form their expectations of future rates based on the general economic outlook, not necessarily only on deviations of the course of monetary policy from the policy rule.

Finally, we explore a more radical departure from the two approaches that we have considered so far, as a way of further investigating the role of shock selection. In particular, we follow the anchoring method used by Altavilla, Giacomini, and Ragusa (2017) in the context of a term structure model, and which goes back to Giacomini and Ragusa (2013) and Robertson et al. (2005). This approach aims to incorporate off-model information in such a way as to minimize the Kullback-Leibler distance between the forecasts with and without the off-model judgement. This methodology remains completely silent about the constellation of shocks that will deliver the desired forecast profile, which provides a robustness check to the more structural approach that we have followed thus far.

When the forecast density is Gaussian,[26] the formulation is very transparent. In particular, if our forecast $y_{t+h|t} \sim \mathcal{N}(\mu_{t+h}, \Sigma_{t+h})$ and our market profile level for the policy rate implies a level of the interest rate R_{t+h} at time $t+h$, our conditional forecast $y_{t+h|(t,R_{t+h})}$ becomes:

$$y_{t+h|(t,R_{t+h})} = \left[y_{t+h|t}[\sim 1] - \Sigma_{t+h}[\sim 1, 1](\Sigma_{t+h}[1, 1])^{-1} \left(y_{t+h|t}[1] - R_{t+h} \right) \right], \quad (11)$$

where we assume without loss of generality that the interest rate is the first entry in the vector y_t, and where the indices in brackets refer to the entries in the vector and matrices (with $[\sim 1]$ meaning all entries but the first).

The formula has a clear parallel to projection theory, and is also economically intuitive. The corrections for the various variables depend on their covariance with the interest rate and the extent to which the off-model information changes the forecast profile for the interest rate. At the extreme, if two variables were orthogonal to each other, conditioning the forecast on off-model information regarding one of them would not affect the forecast of the other at all.

Table 7 reports Diebold-Mariano statistics for the one-year-ahead forecast horizon, comparing the three alternative conditioning scenarios that we considered to the unconditional forecast.[27] As is the case for the results in Table 6, the differences for inflation are not significant, while those for output growth are. In particular, the forecasting performance of the version of COMPASS in which we *anchor* forecasts in the way described in Eq. (11) falls in between those of COMPASS^{R} and COMPASS^{R} (all). The differences are statistically significant and suggest that *anchoring* can improve on imposing the market-rate profile using the monetary policy shock, but is inferior to using the model structure and letting the procedure select the most effective shock constellation. Interestingly, COMPASS^{Anchored} improves on the plain COMPASS forecast for output growth, with the difference being only marginally non-significant at the canonincal levels.[28]

Fig. 17 allows us to investigate the way in which the effects of *anchoring* or conditioning using the model structure changed over our sample. The *anchored* forecast produces results that are generally much closer to those of COMPASS^{R} (all), marginally better when it comes to inflation and somewhat worse when it comes to output growth — which is consistent with the findings of Table 7. Interestingly, Fig. 17 also demonstrates that, while conditioning on the market profile for interest rates is beneficial for forecasting output growth both before and after the crisis (except for the case in which only the monetary policy shock is used for imposing the off-model judgement), the same is not true for inflation. For inflation, conditioning on market-implied rates improved the forecast accuracy prior to the 2008 crisis but not subsequently, which may explain why it is harder to identify significant differences in the forecasting performances of these models when it

[26] Which is the case in COMPASS if we abstract from posterior parameter uncertainty.

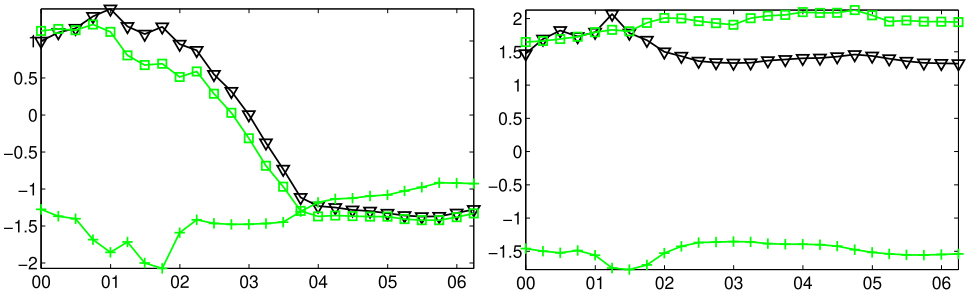
[27] Note that the test statistics for the model pairs that are included in both Tables 6 and 7, namely COMPASS and COMPASS^{R}, differ slightly. This is because, for the purpose of this last exercise, we do not consider posterior uncertainty, meaning that the Gaussian formulas apply exactly. However, the figures are remarkably close, which suggests that shock uncertainty dwarfs the effects of parameter uncertainty in COMPASS.

[28] The implied p-value is around 13%.

Table 7
The role of off-model information: Diebold–Mariano test of equal relative forecasting abilities at a horizon of one year when considering the *anchoring* of forecasts.

	COMPASS	COMPASS^{Anchored}	COMPASS^{R}	COMPASS^{R} (all)
(a) Inflation				
COMPASS	.	−0.733	−1.128	−0.786
COMPASS^{Anchored}	.	.	−0.189	−0.789
COMPASS^{R}	.	.	.	0.059
COMPASS^{R} (all)	.	.	.	.
(b) GDP growth				
COMPASS	.	1.518	−1.780*	2.321**
COMPASS^{Anchored}	.	.	−1.736*	2.159**
COMPASS^{R}	.	.	.	2.032**
COMPASS^{R} (all)	.	.	.	.

Notes: The table shows test statistics. A Newey–West estimator with an optimally chosen bandwidth is used to estimate the long-run variance. If the optimal bandwidth exceeds $2(h-1)$, $2(h-1)$ is used instead. ** (*) indicates statistical significance at the 5% (10%) level. A value that is smaller (larger) than zero indicates that the model in the corresponding row generates more (less) accurate forecasts than that listed in the corresponding column.



(a) Inflation (b) GDP growth

Fig. 17. The role of off-model information: Fluctuation test statistics for a forecast horizon of one year. *Notes:* The window length is seven years. The x-axis shows the start of the window. All models are relative to COMPASS, where a positive (positive) number means that COMPASS performs worse (better). COMPASS^{R} in green with crosses (squares) when using the monetary policy shock (all shocks), COMPASS^{Anchored} is in black with triangles. Critical values for a two-sided test are 2.779 (2.5) at the 5% (10%) levels (Table 1 of Giacomini and Rossi, 2010). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

comes to inflation when we consider the entire sample (as in Table 7).

Thus, the main conclusion of this comparison of alternative approaches to the application of off-model judgment is that using the structure of the model while allowing for all shocks to be used when superimposing off-model judgement is hard to improve upon in this particular example. This approach seems superior to that in which a specific shock is selected, and on a par with, if not superior to, the more "hands-off" *anchoring* methodology.

7. Conclusions

Both accurate forecasts and a structural interpretation of the macroeconomic outlook are important for monetary policymaking. Thus, it is of interest to ask whether different types of off-model information can improve the forecast performances of structural (DSGE) models as a way of retaining a structural interpretation of the forecast, while taking advantage of the flexibility and timeliness of market- or survey-based indicators. In addition, we benchmark the DSGE model's forecast accuracy against statistical and judgmental benchmarks, paying particular attention to instabilities due to the Great Recession.

The accuracies of all forecasts fell during the financial crisis. At the peak of the crisis, performance was poor by recent historical standards, and even in the years that followed, the *Inflation Report* and our model-based methods tended to over-predict GDP growth and under-predict inflation. This deterioration was particularly marked for GDP growth forecasts from the DSGE model. However, the accuracy of the DSGE model's forecasts for GDP can be improved significantly if its information set is augmented with more timely, forward-looking information, in the form of survey expectations of GDP growth. The performances of GDP forecasts from this alternative model are on a par with those of the Statistical Suite and the *Inflation Report*. This is particularly important for improving the accuracy of the model's forecasts during the financial crisis, when timely, forward-looking information was likely to have been particularly valuable.

From a forecasting perspective, these results highlight both the difficulty of dealing with possible structural breaks in real time and the importance of the choice of observable data that are used to condition forecasts from

DSGE models. As it turns out, including measures of survey-based expectations in a DSGE model can at least mitigate some of the adverse effects of suspected structural breaks.

Acknowledgments

A previous version of this paper was circulated under the title "Evaluating UK point and density forecasts from an estimated DSGE model: the role of off-model information over the financial crisis". We thank Charlie Bean, Andy Blake, Marco Del Negro, Raffaella Giacomini, Abi Haddow, Richard Harrison, Simon Hayes, George Kapetanios, Mike McCracken, James Mitchell, Haroon Mumtaz, Gabor Pinter, Simon Price, Kate Reinold, Barbara Rossi, Tatevik Sekhposyan, Kostas Theodoridis, Martin Weale, anonymous referees, seminar participants at the University of Tokyo, the Bank of Japan, Bank of Korea and Banco Central do Brasil, and participants at the 2014 CMEF Conference, the 2014 EMF meeting, the 2014 EABCN Conference on Judgement and Combination in Forecasting and Policy Models, the European Winter Meeting of the Econometric Society, the 25th $(EC)^2$ conference in Barcelona, the 2nd IAAE Conference, 2015 EEA conference and the 2015 CFE conference for their helpful comments and suggestions. We are very grateful to Stephen Burgess and Maddie Warwick for their help in compiling the real-time dataset. Lena Boneva gratefully acknowledges financial support from Cusanuswerk, Germany and the ESRC, United Kingdom.

Appendix. Models in the suite of statistical forecasting models

The Suite of Statistical Forecasting Models (see Kapetanios et al., 2008) provides a set of statistical forecasts of inflation and GDP growth. They are designed to be judgment-free, in the sense that they are unconditional forecasts and do not have any economic structure imposed. Instead, they estimate a set of reduced-form relationships between the variables in each model.

The models use quarterly data that have been transformed such that they are weakly stationary. In particular, the GDP growth measure is the quarter-on-quarter growth in UK GDP, and the inflation measure is the quarter-on-quarter growth of the seasonally adjusted consumer price index. Where necessary, seasonal adjustment is performed using the X12 technique.

The models that comprise the suite, as used in this exercise, are listed below. Each model is estimated for both GDP growth and CPI inflation unless otherwise stated.

- An AR(p) model, in which p is the optimal lag length, selected by the Akaike information criterion.
- A random walk, in which the latest observation of quarterly growth in the variable is projected forward.
- STAR: a variant of the AR(p) model, in which the model fluctuates between two autoregressive regimes. The transition between the two is modelled by a logistic process, the parameters of which are also estimated.
- MSAR: a variant of the AR(p) model, in which the model fluctuates between two autoregressive regimes. The probability of being in one regime or the other is estimated via maximum likelihood.
- Univariate factor model: a variant of the AR(p) model, which also includes a set of principal components taken from a group of over 100 weakly stationary macroeconomic variables.
- VAR(p): a five-variable vector autoregressive model, comprising GDP growth, CPI inflation, the Sterling Exchange Rate Index, the change in the three-month LIBOR interest rate, and quarterly oil price growth.
- VARM: a variant of the VAR, augmented with two variables for capturing the growth in money stocks.
- BVAR: a variant of the VAR, estimated using Bayesian techniques.
- BVARM: a variant of the VARM, estimated using Bayesian techniques.
- FAVAR: a variant of the VAR, augmented with the principal components of a group of over 100 weakly stationary macroeconomic variables. These principal components are treated as endogenous variables.

The individual model forecast densities are generated by simulation. The individual model forecasts are then combined to produce one central point forecast and probability distribution. For the exercises in this paper, the weights have been computed following the predictive likelihood model averaging approach presented by Kapetanios, Labhard, and Price (2006), based on a pseudo out-of-sample window of seven years.

References

Aiolfi, M., Capistran, C., & Timmermann, A. (2010). Forecast combinations. CREATES Research Paper 2010–21, School of Economics and Management, University of Aarhus.

Altavilla, C., Giacomini, R., & Ragusa, G. (2017). Anchoring the yield curve using survey expectations. *Journal of Applied Econometrics*, *32*(6), 1055–1068.

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

Barnett, A., Batten, S., Chiu, A., Franklin, J., & Sebastiá-Barriel, M. (2014). The UK productivity puzzle. *Bank of England Quarterly Bulletin*, *54*(2), 114–119.

Benes, J., Binning, A., & Lees, K. (2008). Incorporating judgement with DSGE models. Reserve Bank of New Zealand Discussion Paper Series DP2008/10, Reserve Bank of New Zealand.

Burgess, S., Fernandez-Corugedo, E., Groth, C., Harrison, R., Monti, F., & Theodoridis, K., et al. (2013). The Bank of Englands forecasting platform: COMPASS, MAPS, EASE and the suite of models. Bank of England working paper.

Cervena, M., & Schneider, M. (2014). Short-term forecasting of GDP with a DSGE model augmented by monthly indicators. *International Journal of Forecasting*, *3*, 498–516.

Christiano, L. J., Eichenbaum, M., & Evans, C. L. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of Political Economy*, *113*(1), 1–45.

Christoffel, K., Coenen, G., & Warne, A. (2010). Forecasting with DSGE models. ECB working paper.

Clark, T., & McCracken, M. (2013). Advances in forecast evaluation. In G. Elliott, & A. Timmermann (Eds.), *Handbook of economic forecasting, Vol. 2, Part B* (pp. 1107–1201). Elsevier.

Clark, T. E., & McCracken, M. W. (2014). Evaluating conditional forecasts from vector autore- gressions. SSRN working paper.

Clements, M., & Hendry, D. (1998). *Forecasting economic time series*. Cambridge, UK: Cambridge University Press.
Corradi, V., & Swanson, N. R. (2006). Predictive density evaluation. In *Handbook of economic forecasting* (pp. 197–284). Elsevier.
Del Negro, M., & Schorfheide, F. (2004). Priors from general equilibrium models for VARs. *International Economic Review*, *4*, 643–673.
Del Negro, M., & Schorfheide, F. (2012). DSGE model-based forecasting. SSRN working paper.
Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.
Edge, R., & Gürkaynak, R. (2011). How useful are estimated DSGE model forecasts? Available at SSRN 1810075.
Edge, R. M., Kiley, M. T., & Laforte, J.-P. (2010). A comparison of forecast performance between Federal Reserve staff forecasts, simple reduced-form models, and a DSGE model. *Journal of Applied Econometrics*, *25*(4), 720–754.
Fawcett, N., Koerber, L., Masolo, R., & Waldron, M. (2015). Evaluating UK point and density forecasts from an estimated DSGE model: the role of off-model information over the financial crisis. Bank of England staff working paper.
Ferroni, F. (2011). Trend agnostic one-step estimation of DSGE models. *The BE Journal of Macroeconomics*, *11*(1), 1–36.
Galvão, A. B., Giraitis, L., Kapetanios, G., & Petrova, K. (2016). A time varying DSGE model with financial frictions. *Journal of Empirical Finance*, *38*, 690–716.
Giacomini, R., & Ragusa, G. (2013). Theory-coherent forecasting. *Journal of Econometrics*, *182*, 145–155.
Giacomini, R., & Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, *25*(4), 595–620.
Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*(6), 1545–1578.
Giannone, D., Monti, F., & Reichlin, L. (2016). Exploiting the monthly data flow in structural forecasting. *Journal of Monetary Economics*, *84* 201–215.
Gürkaynak, R., Kisacikoglu, B., & Rossi, B. (2013). Do DSGE models forecast more accurately out-of-sample than VAR models? *Mimeo*.
Haan, W. J. D., & Levin, A. T. (1996). A practitioners guide to robust covariance matrix estimation. NBER working paper.
Hackworth, C., Radia, A., & Roberts, N. (2013). Understanding the MPC's forecast performance since mid-2010. *Bank of England Quarterly Bulletin*, *53*(4), 336–350.
Iversen, J., Laseen, S., Lundvall, H., & Söderström, U. (2016). Real-time forecasting for monetary policy analysis: The case of Sveriges Riksbank. *Mimeo*.
Joyce, M., Tong, M., & Woods, R. (2011). The United Kingdom's quantitative easing policy: design, operation and impact. *Bank of England Quarterly Bulletin*, *51*(3), 200–212.
Kapetanios, G., Labhard, V., & Price, S. (2006). Forecasting using predictive likelihood model averaging. *Economics Letters*, *91*(3), 373–379.
Kapetanios, G., Labhard, V., & Price, S. (2008). Forecast combination and the Bank of England's suite of statistical forecasting models. *Economic Modelling*, *25*(4), 772–792.
Petrova, K. (2013). Real time forecasting in the recent crisis with a medium-sized DSGE model. *Mimeo*.
Petrova, K., Kapetanios, G., Masolo, R., & Waldron, M. (2017). A time varying parameter structural model of the UK economy. Bank of England working papers 677, Bank of England.
Robertson, J. C., Tallman, E. W., & Whiteman, C. H. (2005). Forecasting using relative entropy. *Journal of Money, Credit and Banking*, *37*(3), 383–401.
Smets, F., & Wouters, R. (2003). An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European Economic Association*, *1*(5), 1123–1175.
Wolters, M. H. (2013). Evaluating point and density forecasts of DSGE models. Economics Working Papers 2013-03, Christian-Albrechts-University of Kiel, Department of Economics.

Lena Boneva's research focuses on topics in macroeconomics, forecasting and panel data econometrics. She completed a Ph.D. at the London School of Economics on panel data econometrics with cross-sectional dependence.

Nicholas Fawcett is a Senior Economist at Goldman Sachs. He was previously a Monetary Policy Adviser at the Bank of England, where this research was completed.

Riccardo M. Masolo is a Senior Economist in the Monetary Analysis area of Bank of England and a member of the Centre for Macroeconomics. Prior to joining the Bank, he obtained an M.Sc. in Economics from Bocconi University and a Ph.D. from Northwestern University.

Matt Waldron has worked in a variety of roles in the Monetary Analysis area of the Bank since joining in 2003. The bulk of the first five years of his career were focused on providing briefing to the MPC on household sector issues, including on the interaction between house prices, debt and consumption. Since then, Matt has worked predominantly on model development and forecasting, having worked on a project to build and implement the Banks latest macroeconomic forecasting platform (which included building a DSGE model of the UK economy and IT to produce forecasts and analyse them).

Matt is currently manager of the Strategy Team, responsible for briefing and provid- ing policy advice to the MPC on monetary strategy issues. His research interests include DSGE modelling, forecasting and monetary policy at the Zero Lower Bound.