BANK OF ENGLAND

Staff Working Paper No. 538

Evaluating UK point and density forecasts from an estimated DSGE model: the role of off-model information over the financial crisis

Nicholas Fawcett, Lena Körber, Riccardo M Masolo and Matt Waldron

July 2015

BANK OF ENGLAND

Staff Working Paper No. 538

Evaluating UK point and density forecasts from an estimated DSGE model: the role of off-model information over the financial crisis

Nicholas Fawcett,[1] Lena Körber,[2] Riccardo M Masolo[3] and Matt Waldron[4]

Abstract

This paper investigates the real-time forecast performance of the Bank of England's main DSGE model, COMPASS, before, during and after the financial crisis with reference to statistical and judgemental benchmarks. A general finding is that COMPASS's relative forecast performance improves as the forecast horizon is extended (as does that of the Statistical Suite of forecasting models). The performance of forecasts from all three sources deteriorates substantially following the financial crisis. The deterioration is particularly marked for the DSGE model's GDP forecasts. One possible explanation for that, and a key difference between DSGE models and judgemental forecasts, is that judgemental forecasts are implicitly conditioned on a broader information set, including faster-moving indicators that may be particularly informative when the state of the economy is evolving rapidly, as in periods of financial distress. Consistent with that interpretation, GDP forecasts from a version of the DSGE model augmented to include a survey measure of short-term GDP growth expectations are competitive with the judgemental forecasts at all horizons in the post-crisis period. More generally, a key theme of the paper is that both the type of off-model information and the method used to apply it are key determinants of DSGE model forecast accuracy.

Key words: DSGE models, forecasting, financial crisis.

JEL classification: C53, E12, E17.

(1) Bank of England and Centre for Macroeconomics. Email: nicholas.fawcett@bankofengland.co.uk
(2) Bank of England and London School of Economics. Email: lena.koerber@bankofengland.co.uk
(3) Bank of England and Centre for Macroeconomics. Email: riccardo.masolo@bankofengland.co.uk
(4) Bank of England. Email: matthew.waldron@bankofengland.co.uk

The views expressed in this paper are our own and not necessarily those of the Bank of England or any of its committees. We thank, without implicating, Charlie Bean, Andy Blake, Spencer Dale, Raffaella Giacomini, Abi Haddow, Richard Harrison, Simon Hayes, James Mitchell, Haroon Mumtaz, Gabor Pinter, Simon Price, Kate Reinold, Barbara Rossi, Tatevik Sekhposyan, Kostas Theodoridis, Sebastian Walsh, Martin Weale, an anonymous referee, seminar participants at the University of Tokyo, the Bank of Japan, Bank of Korea and Banco Central do Brasil, 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 and the 2nd IAAE 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 data set. Lena Körber gratefully acknowledges financial support from Cusanuswerk and the ESRC. The first version of this paper appeared on 1 August 2013; this version was finalised on 3 July 2015.

Information on the Bank's working paper series can be found at
www.bankofengland.co.uk/research/Pages/workingpapers/default.aspx

Publications Team, Bank of England, Threadneedle Street, London, EC2R 8AH
Telephone +44 (0)20 7601 4030 Fax +44 (0)20 7601 3298 email publications@bankofengland.co.uk

ISSN 1749-9135 (on-line)

1 Introduction

A key part of the Monetary Policy Committee's (MPC's) policymaking process is the production and publication of macroeconomic forecasts in its *Inflation Report*. In 2011, the Bank of England introduced a new forecasting platform, which consists of a set of tools used by the staff to help support the MPC's discussions (Burgess, Fernandez-Corugedo, Groth, Harrison, Monti, Theodoridis, and Waldron (2013)). Economic models form an important part of that toolkit and a new DSGE model, the *Central Organizing Model for Projection Analysis and Scenario Simulation* (COMPASS) was introduced at the centre of a suite of models to organise the production of the MPC's forecast. This paper investigates the forecast performance of COMPASS before, during and after the financial crisis, both relative to the MPC's judgemental forecasts and to a statistical benchmark from the Bank of England's Suite of Statistical Models, which combines forecasts from several econometric models (Kapetanios, Labhard, and Price (2008)). Since the MPC's judgemental forecasts are conditioned on a broader information set than these other models, we pay particular attention to how additional, off-model information affects the accuracy of COMPASS's forecasts.

In order to compare the model-based forecasts to the MPC's judgemental forecasts, both COMPASS and the Statistical Suite are re-estimated in each quarter between 2000Q1 and 2013Q1 using real-time data that only reflect information that was available at that point in time. This real-time forecasting exercise makes use of an archive of MPC forecasts for inflation and GDP growth and corresponding data vintages for a range of macroeconomic variables stored by the Bank of England since 1997.[1] Each model vintage is used to produce point and density forecasts from COMPASS and the Statistical Suite that correspond to a set of *Inflation Report* forecasts.

Throughout the period under consideration, the MPC's forecasts have been presented in so-called 'fan charts', representing "the MPC's best collective judgement about the most likely paths for inflation and output and the uncertainties around those central projections."[2] These fan charts provide us with judgemental point and density forecasts, whose accuracy we can assess in the same way as the models' forecasts.

Perhaps not surprisingly, across the sample as a whole, we find that there is no unambiguous winner in the sense that none of the three forecasting methods produces more accurate forecasts for both GDP and inflation at all horizons. At horizons of less than one year, the *Inflation Report* point forecasts are the most accurate for both GDP growth and inflation. At longer horizons, COMPASS has the most accurate inflation point forecasts and the Statistical Suite has the most accurate GDP growth point forecasts. Comparing complete probability density forecasts, the *Inflation Report* is the most accurate for both inflation and GDP at short horizons, but not further out.

Not all of these differences are statistically significant; and forecast accuracy in itself is not the only metric by which models should be assessed. The ultimate objective of the Bank's forecast-

[1] The data archive link is: http://www.bankofengland.co.uk/research/Pages/workingpapers/2015/swp538.aspx.
[2] This text appeared in the foreword of each *Inflation Report*.

ing platform is to support MPC discussions, and models can contribute to this by, for example, elucidating economic mechanisms, providing scenario analysis, and estimating the state of the economy (see Burgess et al. (2013), Section 2.2).

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. Monetary policy responded with unprecedented stimulus: the MPC cut Bank Rate to 0.5% and turned to unconventional policies like Quantitative Easing. Against this backdrop, it is not surprising that many model-based and judgemental forecasts were unable to forecast these developments more than one or two quarters ahead. This is true of the forecast methods we consider in this paper. Figure 1 illustrates this, with none correctly predicting the depth of the recession at the onset of the financial crisis, or the sluggish growth and resilient inflation that followed. This echoes the experience of other central banks and the wider literature (e.g. Del Negro and Schorfheide (2012)).

Although all forecast methods perform badly after the financial crisis, the comparative drop in accuracy of COMPASS's GDP forecasts is particularly marked. We explore two possible explanations in this paper. First, additional information not contained in the model, may help produce better forecasts. For example, the *Inflation Report* predictions, unlike other methods, incorporate timely data, and embody expert judgement. To test for this possibility, we augment the set of observable variables on which COMPASS's forecasts are conditioned to include a timely short-run survey measure of GDP growth. Second, the financial crisis represented a large structural break, and these shifts can lead to serious forecast failure in some types of model. Trend productivity growth in the standard version of COMPASS is calibrated using data from the twenty year period up to the financial crisis. But 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 a model not robust to such a shift.[3] With this in mind, we also produce forecasts from a version of COMPASS in which we allow the productivity growth trend to be time-varying.

Both additional information, and a time-varying productivity trend, help to produce more accurate GDP growth forecasts after the crisis. Incorporating survey data on GDP growth leads to a statistically significant improvement in forecast accuracy; and by conditioning on the Bank staff's short-term inflation projection for the first six months of the forecast, COMPASS produces more accurate inflation forecasts up to a year ahead. Both of these results suggest that the application of timely off-model information can materially improve forecasts from DSGE models, particularly during times of macroeconomic volatility. Allowing the productivity trend to vary over time does improve forecast accuracy, but the gain is small (and not significant) compared to incorporating the GDP growth survey data.

Finally, we also assess how the application of the financial conditioning assumptions used in the *Inflation Report* forecasts – that interest rates evolve in line with market expectations and

[3] Clements and Hendry (1998) discuss the robustness of models to breaks in long-run trends.

that the exchange rate follows a path that is an average of that implied by a UIP condition and that implied by a random walk – alters the accuracy of COMPASS's forecasts. The result depends crucially on the assumptions under which the conditioning paths are applied. In particular, if the market interest rate expectations are assumed to contain information about the future state of the economy, then conditioning COMPASS's interest rate forecast to match them improves the accuracy of the inflation forecasts at almost all horizons, and the GDP forecasts at shorter horizons. If, on the other hand, market interest rate expectations are assumed to contain information about the future stance of policy only, then the accuracy of COMPASS's inflation and GDP forecasts deteriorates at almost all horizons.

There is already an extensive literature that has assessed the forecasting performance of DSGE models. The majority of the literature (Del Negro and Schorfheide (2012), Gürkaynak, Kisacikoglu, and Rossi (2013), Edge and Gürkaynak (2011), Wolters (2013)) has investigated forecasts from small-scale models, similar to Christiano, Eichenbaum, and Evans (2005) and Smets and Wouters (2003). There is, however, also some work for models of other central banks, including the Swedish Riksbank (RAMSES; Iversen, Laseen, Lundvall, and Söderström (2013)); the ECB (New Area Wide Model, Christoffel, Coenen, and Warne (2010)); and the Federal Reserve Board (EDO, Edge, Kiley, and Laforte (2010)). These studies all differ in the comparator forecasts used to evaluate 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. (2013) do not. In addition, there is substantial variation in the statistical methods used to evaluate the forecasting performance of the DSGE models. While all papers report root mean squared forecast errors, only some perform statistical tests to assess if the forecasting performance of alternative models is significantly different (Gürkaynak et al. (2013), Edge and Gürkaynak (2011), Edge et al. (2010)). We formally test if the accuracy of alternative forecasts is significantly different and we also investigate instabilities in forecast performance.

The remainder of this paper is organised 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 to evaluate alternative forecasts. Section 5 reports results on the accuracy of alternative point and density forecasts. Section 6 discusses how the forecast performance is affected by imposing alternative sets of conditioning paths. 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 containing, among other things, the MPC's best collective judgement density forecasts for inflation and GDP growth. In 2011, the Bank of England introduced a new forecasting platform to assist with the production of those *Inflation Report* forecasts (Burgess et al. (2013)). This paper evaluates the performance of forecasts from two key models that form part of that

platform with reference to the performance 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 best supporting the process of producing the MPC's forecasts.[4] Given the judgemental nature of those forecasts, this process puts a premium on discussion of the economics of the forecast, rather than a desire to maximise forecast accuracy of all of the models used. Reflecting that, 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 organising device in the construction of the MPC's judgemental forecasts.[5] Since it is built on economic theory, COMPASS 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,[6] and COMPASS is no exception, so a key design feature of the forecasting platform is that it recognises the importance of a suite of models. The suite contains a large number of different models of varying types and classes, each with different purposes in mind.[7] These produce alternative forecasts to cross-check the MPC's forecast. And, 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 explicitly designed with forecasting performance in mind. The rest of this section provides some brief background on COMPASS, the Statistical Suite and the *Inflation Report* forecasts, which we evaluate in Section 5.

2.1 COMPASS

COMPASS is a medium to large-scale New Keynesian DSGE model built on the tradition of Smets and Wouters (2003) and Christiano et al. (2005), and with similarities to antecedents at other central banks. Like all models of its class, monetary policy can influence the paths of activity and inflation in the short to medium term because prices are assumed to be sticky. The model economy is populated by households, firms, a central bank, a government, and an exogenous rest-of-the-world economy. Figure 2 summarises the interactions between those agents, where boxes represent sectors and arrows illustrate the flow of goods and services between them. For a detailed derivation and description of the model equations, we refer readers to Burgess et al. (2013).

COMPASS is estimated with Bayesian maximum likelihood methods on UK data for 15 macro-

[4] See Burgess et al. (2013), Section 2 for a much fuller discussion.

[5] As its name suggests, COMPASS is also used as a laboratory to quantify the macroeconomic effects of various shocks (scenario analysis) and alternative assumptions about policy (policy analysis) – see Section 7 of Burgess et al. (2013) for examples.

[6] This is often framed in terms of misspecification. There are two related ways in which a model might be misspecified. First, a model can be statistically misspecified if the assumptions underpinning it are violated by the data (e.g. if residuals in an OLS regression are not *iid*). Second, a model can be economically misspecified (and all models are simplifications of reality by definition) if it does not articulate an economic shock or transmission mechanism that is of relevance given its use (e.g. COMPASS does not contain an energy sector and energy supply (or price) shocks drive at least some of the variance of CPI inflation).

[7] See Burgess et al. (2013), Section 5 for discussion of the suite of models.

economic time series[8] using 18 shocks[9]. One of those shocks is a permanent labour augmented 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. As discussed in Section 5.5, this shock has an important role to play in shaping the forecast performance of COMPASS in the post-crisis period. 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).[10] The second group of parameters are estimated using Bayesian maximum likelihood and mainly include parameters that govern the model's dynamics (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 described in detail in Section 4.3 of Burgess et al. (2013).[11]

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 were not tackled in that paper. In particular, the MPC cut Bank Rate to its effective lower bound of 0.5% and implemented a Quantitative Easing (QE) programme. Given 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 policy rate observable rather than Bank Rate as in Burgess et al. (2013).[12] This shadow measure augments Bank Rate to include an in-house estimate of the effect of QE.[13]

In order to investigate the importance of timely off-model information and trend specification in determining the forecast performance of DSGE models, Section 5 evaluates three alternative versions of COMPASS against forecasts from the Statistical Suite and the *Inflation Report*:[14]

1. The Plain variant of COMPASS (COMPASS^{PLAIN}) has a structure described by Burgess et al. (2013), but uses the shadow measure of the policy rate described above (which is only rele-

[8]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, 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.

[9]These are detailed in Appendix A.

[10]As described in Section 4.3 of Burgess et al. (2013), these parameters are typically calibrated to match sample averages.

[11]In particular, we use the same prior distributions and an estimation sample beginning in 1993Q1 with data from 1987Q2 to 1992Q4 used as a training sample for the Kalman filter.

[12]The extended estimation period also covers several changes to VAT. We therefore measure CPI inflation excluding the effects of these changes. Burgess et al. (2013) Section 8.2 describe how Bank staff estimate the effect of VAT changes on inflation.

[13]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 underlying assumption that underpins this approach is that QE is a close substitute as a monetary policy instrument to Bank Rate such that the zero lower bound was not an effective constraint on monetary policy over the period in question.

[14]The value of imposing other conditioning paths like a market measure of interest rate expectations is assessed in Section 6.

vant after 2009Q1) as the policy rate "observable".

2. The version with Time-Varying-Trend (COMPASS^{TVT}) is based on the Plain variant, but uses a productivity trend computed as a seven year moving average, to capture potential shifts in the trend growth rate of the economy. This can be motivated in several ways. For example, Comin and Gertler (2006) document that the the growth rates of developed countries tend to oscillate at medium-term frequencies between persistent periods of robust growth and relative weakness. Allowing the productivity trend to change over time could capture this phenomenon. Moroever, calculating trends over a rolling window is more robust to instabilities and structural breaks that are frequently encountered in macroeconomic time series (and so this is very much in the spirit of the estimation of the Statistical Suite described below). The choice of the productivity trend for the rolling calculation can be motivated by the extraordinary productivity outturns observed since the crisis (see, e.g. Barnett et al. (2014)).[15]

3. The Survey-augmented version (COMPASS$^{GDP^e}$) adds a survey measure of one-year ahead growth expectations to the set of observables used to identify the state in the Plain model.[16] Because COMPASS^{PLAIN} contains more shocks than observables[17], this does not require adding new shocks to the model. COMPASS$^{GDP^e}$ uses the parameter estimates from COMPASS^{PLAIN}.[18] Note that as discussed in Section 3.1, 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.

2.2 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 are designed specifically to produce forecasts of inflation and GDP growth. As such, they do not have a particular economic structure in the way that COMPASS does, and instead they estimate a set of 'reduced-form' relationships between macroeconomic variables.

The Suite therefore has the value of being 'judgement free'. It is not constrained by any theoretical restrictions, unlike COMPASS; and it does not incorporate expert judgement from the MPC, unlike the *Inflation Report* forecasts. Furthermore, the forecasts are not conditioned on any path for the exchange rate or monetary policy.

[15] As such, there is a question of whether or not this would have been something that would reasonably have been implemented in real time.

[16] 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 appears in each *Inflation Report*.

[17] See Burgess et al. (2013, Section 4.3.2) for a discussion.

[18] That is, we do not re-estimate the model with the additional observable. While this would be desirable in principle, we choose not to given the cost of the real-time estimation.

It includes both univariate and multivariate forecasting methods. The set of univariate models includes random walk, 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 all are estimated over a seven year rolling window. This provides a degree of robustness to structural change. The forecasts obtained from 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 C.

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

2.3 *Inflation Report*

Since monetary policy independence in 1997, the Bank's quarterly *Inflation Report* has communicated the MPC's assessment of the economic outlook. As part of that assessment, the MPC produce 'fan charts', which represent their best collective judgement about the most likely paths for inflation and GDP growth and the uncertainty surrounding them.[19] These fan charts provide us with a sequence of real-time judgemental density forecasts for inflation and GDP growth against which we can compare the model-based forecasts.

In making this comparison two points stand out. First, by convention the *Inflation Report* forecasts take as given the paths of several several variables, including the government's announced fiscal plans, and a particular path for monetary policy.[20] Under the assumption that the forecasts we evaluate are 'primitives', the statistical techniques we use are valid, but the issue should be borne in mind when comparing the *Inflation Report* forecasts with the unconditional model-based forecasts.[21] In Section 6, we explore the impact of applying the same conditioning information to the DSGE model's 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. Models are used by the staff to aid the deliberations of the MPC, but the final published forecast represents the Committee's best collective judgement. The forecast performance of one does not therefore determine that of the other.

[19] Technically speaking, the density forecasts are constructed from two separate two-part normal distributions in which the MPC decide the mean, variance and skew.

[20] 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.

[21] The comparison would be unaffected by this issue if the conditioning assumptions on which the *Inflation Report* forecasts are made had no systematic impact on the judgemental projections that were published.

3 A real-time dataset for estimation and forecast evaluation

This section describes the real-time dataset used for estimation of the models and evaluation of forecast accuracy. Throughout, our objective is to evaluate the models' forecast accuracy using only information and data that were available at the time the corresponding *Inflation Report* forecast was made.[22] In addition, in order to make the lessons of this exercise practically applicable, we also have the objective of estimating the models and constructing forecasts in a way that is broadly consistent with their use in the Bank.

3.1 Estimation data

Our dataset is constructed from a set of real-time forecasts and associated datasets that have been stored by the Bank of England since 1997.[23] For most of the variables required to estimate the two models, the real-time data can be taken directly from this database. For others, however, there had been substantial definitional changes and/or the required series had not been stored in the database over some of the forecast evaluation sample.[24] In such cases, we manually reconstructed series in a way that was consistent with our objective of using only information that was available prior to each forecast origin.[25]

Our evaluation exercise covers the period 2000Q1-2013Q1. As discussed in Section 2, we re-estimate COMPASS over this period using a recursive scheme with an estimation sample beginning 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 both models are 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 we use to evaluate 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 with a dataset constructed approximately one month ahead of each *Report*.[26] Figure 3 illustrates the timing of data

[22] There are limits to the extent to which our exercise can be regarded as truly 'real time'. For example, although the dataset 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.

[23] The variables required by COMPASS are similar to those required by previous organising models. Prior to introducing COMPASS, the central organising model was the Bank of England Quarterly Model (BEQM) (Harrison, Nikolov, Quinn, Scott, and Thomas (2005)), and prior to that, it was the Medium Term Macroeconomic Model (MTMM) (Bank of England (2000)). Alongside the publication of this paper, we have also published the full archive of real-time datasets that we constructed. This is available at http://www.bankofengland.co.uk/research/Pages/workingpapers/2015/swp538.aspx.

[24] For example, the measure of wages used in BEQM was wages per head in the private sector, whereas they are defined as a whole economy measure in COMPASS. This meant that we had to reconstruct a whole economy measure using the private sector measure and a measure of public-sector wages that was part of the BEQM-based database.

[25] Details on the construction of the real-time dataset are available from the authors upon request.

[26] This is true for the February, May and November *Inflation Reports*. The time lag is around three weeks in the case

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 in order to mimic the real-life conditions in which the models are used, we maintain this convention in our evaluation.

We also condition the real-time forecasts with "nowcasts", or near-term forecasts, which are based on Bank staff's judgement and statistical models. For each *Inflation Report*, the nowcast quarter is the one prior to the quarter in which publication occurs. For example, the nowcasting quarter for the November 2014 *Inflation Report* was 2014Q3 (Figure 3). At Benchmark stage, some data 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 nowcasts for the missing variables (which are released for the first time between Benchmark and the *Inflation Report*). The nowcasting quarter 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 the nowcasting quarter, but the data used in the real-time estimation of the Statistical Suite do not. For example, in using COMPASS to produce a forecast corresponding to the November 2012 *Inflation Report*, we use real-time data between 1993Q1 and 2012Q2 to estimate COMPASS and real-time data between 1999Q4 and 2012Q3 (incorporating nowcasts) to estimate the Statistical Suite (reflecting the seven year rolling window). In order to ensure that the information sets used to produce the forecasts are the same, we impose the nowcasts as judgement in the first quarter of forecasts constructed with COMPASS (using all of COMPASS's shocks in a "full inversion").[27]

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 that are published with a lag of about three months. An alternative would have been to use the preliminary release that is available with a lag of two months. In the forecast evaluation exercise of Del Negro and Schorfheide (2012), these two approaches give very similar results.[28]

One complication in evaluating the inflation forecasts is that the MPC's target measure of inflation changed in December 2003 from the RPIX to the CPI series. So 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 inflation forecasts from the Statistical Suite and COMPASS.[29]

of the August *Inflation Report*.

[27] In Section 6 we also explore the impact of conditioning COMPASS's forecasts on the staff's short-term inflation forecast, which is used in the construction of the MPC's *Inflation Report* forecasts.

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

[29] In our evaluation, COMPASS forecasts CPI inflation excluding the impact of VAT changes; we therefore compare these projections to an estimate of CPI inflation excluding VAT. This does not affect the conclusions of our evaluation.

4 Statistical methods to evaluate forecast accuracy

Before describing the statistical tests, we first introduce some notation. At each forecast origin t, we observe the data $\{y_j^t, x_j^t\}_{j=1}^s$, where $s \leq t$, y is a vector of variables to be predicted and x is a vector of additional variables that are included in the forecasting model. In our data it is typically the case that $s < t$ because many data are published with a lag. The dependence of y_j and x_j on the forecast origin t reflects that this is a real-time forecasting exercise and that some data get revised over time to incorporate new information or measurement methodologies.

We split the sample, which is of length T, into an in-sample period of length R and an out-of-sample period of length P; h is the forecast horizon. This leaves us with a sequence of out-of-sample forecasts from $R+h$ to T, to evaluate; or equivalently, we have $P-h+1$ forecasts.[30] The time point of the forecast origin is denoted t; the first origin is at point R, and the most recent is at $T-h$.

Throughout the paper, the estimation error is captured under the null hypothesis which means that we adopt the asymptotic framework of Giacomini and White (2006) to conduct inference. That framework is applicable if the parameters are estimated using a small rolling window of observations. The Statistical Suite fits well in this framework as it is estimated over a rolling window of 28 observations. Strictly speaking, this test is not directly applicable to the recursive scheme used in the real-time re-estimation of COMPASS. However, under the assumption that COMPASS (and *Inflation Report*) forecasts are 'primitives,' the estimation scheme should not matter.[31]

A limitation of the statistical methods we use is that they do not take into account the real time nature of our data. Depending on the properties of the data revisions, statistical tests to evaluate forecast accuracy can have non-standard asymptotic properties (Clark and McCracken (2013)). This question is left for future research.

4.1 Accuracy of point forecasts

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

$$RMSFE^h = \sqrt{\frac{1}{P-h+1}\sum_{t=R}^{T-h}\widehat{u}_{t+h}^2} \tag{1}$$

where the forecast error $\widehat{u}_{t+h}$ at forecast origin t and forecast horizon h is defined as the difference between the data and the mean forecast. It depends on both parameter estimates at forecast origin t and the outturn of the variable being forecast.

[30] Since our evaluation period runs from 2000Q1 to 2013Q1, the out-of-sample window is 53 quarters; so there are $54-h$ forecasts (h steps ahead) to evaluate.

[31] In a simulation study, Clark and McCracken (2013) document that the Giacomini-White test works well for both rolling and recursive schemes.

To test if RMSFEs from alternative forecasts are significantly different, Diebold and Mariano (1995) proposed the following test statistic:

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

where $\widehat{\Sigma}$ is an estimate of the long-run variance. We use the Newey-West estimator where the bandwidth is chosen optimally. Because of the slow rate of convergence of heteroskedasticity and autocorrelation consistent covariance estimators, they can perform poorly in small samples (Haan and Levin (1996)). To address this concern, Appendix B documents the small sample properties of the Diebold-Mariano test. We find that in sample sizes like ours, the Diebold-Mariano test has correct size and acceptable power.

Under the null hypothesis $H_0 : E(\widehat{u}_{1,t+h}^2 - \widehat{u}_{2,t+h}^2) = 0$, DM_h converges to a normal distribution, provided that the loss difference is covariance stationary and has a constant mean and variance.[32]

The Diebold-Mariano test is suitable to assess the *unconditional* relative predictive ability of two alternative forecasting models. But it is also interesting to investigate if, for example, one model predicts more accurately in times of high uncertainty while another model performs better in normal times. To this objective, Giacomini and White (2006) propose the following regression model:

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

where X_t contains information that is known at the forecast origin t such as a constant, indicators of economic activity or measures of global uncertainty. If X_t contains only a constant, the Giacomini and White (2006) test is equal to the test of Diebold and Mariano (1995). Under the null hypothesis $H_0 : \widehat{\alpha} + \widehat{\beta}E(X) = 0$, two alternative point forecasts are equally accurate *conditional* on X_t.

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

$$GW_h = (P-h+1)\overline{Z}'\widehat{\Sigma}^{-1}\overline{Z} \tag{4}$$

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

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

and $\widehat{\Sigma}$ is an estimate of the long-run variance $\Sigma = \lim_{P\to\infty} V((P-h+1)^{1/2}\overline{Z})$. To estimate Σ, we use a Newey-West estimator where the bandwidth is chosen optimally. Asymptotically, the Giacomini-White test has a $\chi^2(2)$ distribution.

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

If H_0 is rejected, then it is possible to predict which method has 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 to use the first model 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 unrealistic in unstable environments.

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

4.2 Accuracy of density forecasts

While it is straightforward to compare a point forecast to the actual outcome, it is more difficult to evaluate a complete density forecast. This is because we never observe the entire distribution, but rather, only one realisation from it. A popular approach to overcome this constraint is to calculate how likely it would be under the forecast density to observe the realised value. For example, inflation was 3.5% in 2012Q1 and the February 2012 *Inflation Report* predicted a 60% chance of achieving this, or a lower inflation rate. 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 realisation 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 under the forecast distribution of observing a value that is equal to or less than the realised value.

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

$$\begin{aligned} P(F(Y) \leq z) &= P(Y \leq F^{-1}(z)) \\ &= F(F^{-1}(z)) \\ &= z \end{aligned} \tag{7}$$

Given the small sample of forecasts that we have available, tests of uniformity in the distribution of PITs $\{z_s\}_{s=1}^{T}$ are generally unsuitable. As Berkowitz (2001) notes, many non-parametric

tests are notoriously data-intensive, suggesting the need for at least 1,000 observations for a Kolmogorov-Smirnoff test of uniformity. We therefore follow the approach suggested by Berkowitz (2001), in which we transform the PITs to create a new variable $z_t^* = \Phi^{-1}(z_t)$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal CDF. This does not impose any distributional assumptions on the underlying data, so, as Berkowitz (2001, p. 467) notes, "inaccuracies in the density forecast will be preserved in the transformed data." The relevant test is therefore that the transformed PITs have zero mean, and unit variance, which would be the case under the null hypothesis that the PITs themselves were uniformly distributed over the unit interval. We follow Berkowitz (2001) by estimating an AR(1) model for z_t^*, and perform a $\chi^2(2)$-distributed likelihood ratio test for zero mean and unit variance in the z_t^*s.

A related method to evaluate 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 follow Alessandri and Mumtaz (2013) and 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 by the average logarithmic score that is defined as:

$$\frac{1}{P-h+1}\sum_{t=R}^{T-h}\log f_t(y_{t+h}) \tag{8}$$

To test whether logarithmic scores from competing models are significantly different, we apply the (unweighted) likelihood ratio test of Amisano and Giacomini (2007).[33] Under the null hypothesis $H_0 : E(\log f_{1,t}(y_{t+h}) - \log(f_{2,t}(y_{t+h})) = 0$, the two density forecasts $f_{1,t}(.)$ and $f_{2,t}(.)$ perform equally well. The dependence of the forecast densities on past data and on the estimated parameters is suppressed. The test statistic is given by:

$$AG_h = \frac{1}{\sqrt{P-h+1}}\sum_{t=R}^{T-h}\frac{\log f_{1,t}(y_{t+h}) - \log f_{2,t}(y_{t+h})}{\sqrt{\widehat{\Sigma}}} \tag{9}$$

where $\widehat{\Sigma}$ is an estimate of the long-run variance that is estimated in the same way as above. Asymptotically, AG_h is distributed as a $N(0,1)$ variable.

4.3 Forecast evaluation in the presence of instabilities

The Diebold-Mariano and logarithmic score tests assess relative forecast performance on average over the sample under consideration. But relative performance can change over time, as certain events – such as the financial crisis – affect the predictive content of the variables be-

[33] A more general version of the Amisano-Giacomini test is developed in Giacomini and White (2006). In future work, we could also consider a conditional version of the test by Amisano and Giacomini (2007) and a time-varying version that is inspired by the fluctuation test.

ing 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.

To test the null hypothesis that both models have equal predictive ability at each point in time, or $H_0 : E(u_{1t}^2 - u_{2t}^2) = 0 \ \forall \ t$, Giacomini and Rossi (2010) propose the fluctuation test statistic:

$$F_{P,h} = max_t|F_{t,h}| \tag{10}$$

where:

$$F_{t,h} = \frac{1}{\sqrt{m}} \sum_{j=t-m/2}^{t+m/2-1} \frac{\log f_{1,t}(y_{t+h}) - \log f_{2,t}(y_{t+h})}{\sqrt{\widehat{\Sigma}}}, \quad t = R + \frac{m}{2}, ..., T - \frac{m}{2} + 1 \tag{11}$$

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 in 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 then the second model has superior forecasting performance, while the first model is preferred if the lower bound is crossed.

5 Evaluation of point and density forecasts

5.1 Descriptive results

We start by graphically comparing the forecasts from the three alternative versions of COMPASS, the Statistical Suite, and the *Inflation Report* against actual outcomes (Figures 4–8).

The inflation projections shown in the left-hand panels of each figure tend to be attracted towards an 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 leading up to the forecast origin. This explains why, as the actual inflation rate tended to rise over the evaluation window, the suite inflation forecasts for one to two years ahead were progressively higher. In the case of COMPASS variants, the path of inflation over the forecast horizon is partly driven by a natural mean-reversion back to the inflation target of 2%. This is most evident in Figure 6a. But in this larger model, the deviation of inflation from 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, in the Plain and Time-Varying-Trend versions of COMPASS, as shown in Figures 4a and 5a, since the forecast for GDP growth overshoots the trend, as discussed below. There is no automatic mean-reversion of the *Inflation Report* forecasts, as the projections reflected the MPC's judgement about the impact of the state of the economy, and monetary policy, at the time of the forecast. But while the Committee's forecasts of inflation proved to be quite accurate before the crisis, the post-crisis period saw systematic underpredic-

tion.

All GDP growth forecasts overshoot the outturns at some point in the evaluation period. This is particularly marked for the Plain variant of COMPASS in Figure 4b; its tendency to over-predict growth is accentuated after the crisis, both in comparison to earlier performance and to the other forecast methods considered in this paper. To understand why this happens, and draw out some comparisons of the different forecast methods, we consider the role of two drivers. First, the underlying long-run growth rate that underpins each forecast model; and secondly, additional information and judgement that is incorporated into the *Inflation Report* projections, but which is not necessarily incorporated into forecast models.

In all variants of COMPASS and the Statistical Suite, the speed with which 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 an equilibrium growth rate (comprising 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, although the exact speed of convergence will depend on the nature of the underlying shocks at the forecast origin.[34] The accuracy of these GDP growth forecasts therefore depends on whether the estimate of long-run growth is correct. In the Plain variant of COMPASS, trend growth is calibrated using data over the fifteen years leading up to the financial crisis, and is not assumed to have changed since then. Against a backdrop of sustained weakness over the past six years, it is not surprising that the model would therefore significantly overpredict growth.[35] Models whose equilibrium growth rates can vary over time, are likely to be more robust to the kind of shocks wrought by the financial crisis. Both the Statistical Suite and the variant of COMPASS with a Time-Varying Trend, incorporate trends estimated over a seven-year period running up to the forecast origin. This improves forecast performance, especially after the crisis, as can be seen in Figures 7b and 5b respectively. With each successive forecast origin after 2008, the estimated trend growth rate falls as the crisis begins to dominate the sample period, in turn pushing down on the GDP growth forecasts. This result chimes with other papers in the literature that stress the importance of correctly specifying trends 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 longer – or shorter – to return to its trend rate, than would normally be the case. This alludes to the role of expert judgement, and additional information not normally incorporated into a forecast model, in forecasting. The

[34]Common to 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. Those sample averages tend to exceed the growth rates observed in the data after the crisis (and this is something that the rolling productivity variant mitigates). 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 and so there is some mean reversion built into the level of GDP in COMPASS, as well as the growth rate. Section 5.5 discusses the shocks identified by COMPASS over the crisis period in a bit more detail.

[35]Oulton and Sebastiá-Barriel (2013) suggest that, while the financial crisis may not have permanently reduced the UK's long-run growth rate, it has had a long-lasting effect on the level of GDP.

Inflation Report forecasts in Figure 8b illustrate the former. In the immediate aftermath of the crisis, the MPC predicted a rapid recovery in GDP growth rates, which failed to materialise. As Hackworth, Radia, and Roberts (2013) note, this forecast failure broadly reflected three judgements of the Committee that turned out to be wrong: 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. But although in this instance, the additional information embodied in these judgements led to worse forecasts, in general it could be beneficial. Figure 6b shows that the variant of COMPASS incorporating survey expectations of GDP growth, produces better GDP growth forecasts than the Plain version, particularly after the crisis. In this case, the additional information leads COMPASS to appeal to a different mix of underlying shocks to explain the evolution of the financial crisis, compared to the Plain variant. It finds that the data are explained better by more persistent shocks – including a permanent shock to productivity – which depress GDP growth for longer.

The broad point from this and the *Inflation Report* forecasts is therefore that a deeper understanding of the factors pushing GDP growth away from its equilibrium can sometimes be of help when forecasting. This applies in particular to information that is relevant during the financial crisis. Del Negro and Schorfheide (2012) demonstrate that a version of the Smets-Wouters model modified to incorporate financial frictions, following Bernanke, Gertler, and Gilchrist (1999), forecasts GDP growth more accurately during the financial crisis than the equivalent model without these frictions. It is possible that the survey expectations measure we use is picking up some of the same information as the measure of credit spreads used by Del Negro and Schorfheide (2012) in their model. Moreover, they show that the model's forecasting performance over the crisis period can be improved still further, by conditioning the forecasts on the latest available measure of spreads. This suggests that the forecast performance of the Survey-augmented variant of COMPASS could also be improved if we were to condition it on the *contemporaneous* survey expectation instead of the *lagged* measure.

5.2 Point forecast evaluation

To evaluate the point forecast accuracy of each of the alternative forecast methods, we use 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. We then investigate whether these differences were predictable given the information available at the time each forecast was made. In this section, we concentrate on performance across the sample as a whole; Section 5.4 assesses whether forecast performance changes over time.

The RMSFEs for all forecast methods are shown in Figures 9a and 9b for inflation and GDP growth respectively. Looking first at the near-term forecasts (between three to five quarters

ahead), the *Inflation Report* forecasts are more accurate than any model-based forecasts.[36] Further ahead the rankings reverse, in particular for inflation, where the RMSFE of *Inflation Report* forecasts is highest of all methods. Echoing the discussion above about the benefits of incorporating additional information into projections, the forecasts from the Survey-augmented variant of COMPASS are more accurate than the two alternative variants. But of all the differences in forecast accuracy, only this last one is statistically significant, when tested using the method of Diebold and Mariano (1995). As shown in Tables 1 and 2, the forecasts from the Plain and Time-Varying-Trend versions of COMPASS are significantly less accurate over the sample than those of the *Inflation Report*, Statistical Suite and Survey-augmented COMPASS at a one-year horizon.

Statistical significance notwithstanding, the finding that judgemental forecasts tend to perform better at short horizons and model-based forecasts better at longer horizons is consistent with many papers in the literature. Del Negro and Schorfheide (2012) find that forecasts for inflation and GDP growth from the Smets-Wouters model are more accurate than the Federal Reserve Board staff's judgemental Greenbook forecasts at horizons of three quarters or longer; Edge et al. (2010) find the 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* in forecasting inflation at some horizons. In contrast, Iversen et al. (2013) document that the Monetary Policy Committee of the Riksbank tends to outperform forecasts obtained from their DSGE model RAMSES or from a Bayesian vector auto-regression.

The Diebold-Mariano test assesses the *unconditional* relative predictive ability of two alternative forecasting models. But it is also interesting to investigate if the relative performance varies *conditional* on information that was available at the forecast origin. Motivated by the substantial tightening in credit conditions over the financial crisis and the volatility of oil prices between 2007–2009, 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 and two year horizons. Tables 3–4 report Giacomini and White (2006) test statistics.[37] We find that both variables can predict loss differences for GDP growth one year ahead in several of the comparisons.[38] This suggests that, conditional on the credit spread, the Statistical Suite is more accurate than both Plain and Time-Varying Trend versions of COMPASS, but not the variant incorporating the survey measure of GDP expectations. Given that COMPASS does not include financial frictions, one hypothesis for this is that these survey expectations, the Statistical Suite and *Inflation Report*, implicitly incorporate information about credit conditions that is not available to the Plain COMPASS model. Compared to GDP growth, forecasting loss differences in inflation is more difficult (Tables 3 and 4).

[36] Model-based methods comprise all variants of COMPASS, and the Statistical Suite.

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

[38] Although not reported here, house prices, the VIX and the sterling ERI can predict forecast accuracy differences in GDP growth as well.

5.3 Density forecast evaluation

Probability Integral Transforms (PITs) are a useful method to assess whether or not a density forecast is consistent with the observed frequency of outturns. Figures 10 and 11 illustrate the PITs for inflation and GDP growth respectively for all five of our forecasting methods. Each blue dot shows the probability – under the forecast distribution – of observing an outturn equal to or lower than the inflation or GDP growth rate that came to pass. The size of blue dots is proportional to the frequency of observing a given probability score. If the density forecasts were accurate, we would expect the dots to be roughly evenly spaced out between 0 and 1.

At shorter horizons, the Plain and Time-Varying-Trend versions of COMPASS tend to underpredict inflation (Figure 10), as illustrated by the concentration of observations in the the upper tail of the distribution. Furthermore, at short forecast horizons and for all COMPASS variants, too many PITs are located in either the highest or lowest percentile buckets, indicating that they all understate the uncertainty around the inflation forecast, and the same appears to be true of the Statistical Suite. The one-quarter ahead inflation forecasts in the *Inflation Report* tend to show the opposite problem, however, as they overstate the uncertainty around the point forecast. At longer horizons, the inflation PITs for the Plain and Time-Varying-Trend versions of COMPASS, and the Statistical Suite seem to be more evenly spaced between 0 and 1, while there is evidence that the Survey-augmented COMPASS and the *Inflation Report* underpredict inflation.[39]

Figure 11 illustrates that GDP growth is overestimated by all five of our forecasting methods, as there are very few outturns in the upper percentile buckets. It is also clear that the Survey-augmented version of COMPASS produces the best GDP forecasts of all three variants. Overall, at longer horizons the GDP density forecasts from the Statistical Suite appear to be the best calibrated with a more even spread of PITs between 0 and 1.

Assessing the PITs more formally, Table 7 reports the results of the Berkowitz (2001) test of the transformed PITs. This is a $\chi^2(2)$ test that the transformed PITs have a zero mean and unit variance, and at both one and two year horizons, all forecast models fail this test. Unfortunately the test does not distinguish between incorrect specification of the mean or the variance as the source of failure. However, given the visual pattern identified above, it is likely to be the result of a mis-specified forecast mean.

Scoring rules are a convenient way of summarising the information 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 realised outturns (Section 4.2). Figure 12 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 of the Statistical Suite are significantly more precise than those of all other methods (Table 5). The inflation forecast density of Survey-augmented

[39]The difference in the inflation density forecast performance of COMPASS$^{GDP^e}$ relative to COMPASS^{PLAIN} reflects the difference in shock identification, as discussed in Section 5.1.

COMPASS is very inaccurate at short horizons, but further out it is not far from the Statistical Suite. For GDP growth, Survey-augmented COMPASS, the *Inflation Report* and the Statistical Suite are more accurate than the other versions of COMPASS at near-term horizons. But only a few of these differences in forecast accuracy are statistically significant at the one-year horizon (Table 6).[40] At longer horizons, the Statistical Suite forecasts are substantially worse than other methods. This is largely due to the financial crisis period, during which some of the GDP growth outturns were a long way into the lower tail of the Suite probability forecast. Given the nature of the logarithmic transformation of these small probabilities, the penalty applied to the Suite is significant over this period.

5.4 Instabilities in forecast performance

In Sections 5.2 and 5.3, we evaluated the performance of forecasts from the three COMPASS variants, the Statistical Suite and the *Inflation Report* across the entire evaluation sample. We now turn to investigate how that performance has varied over time, paying particular attention to the financial crisis as a cause of instability.

Figure 13d reports average logarithmic scores separately for the pre- and post-crisis period. For both inflation and GDP growth, the minimum log score (looking across all five sources) is lower in the earlier period than the later, indicative of a general deterioration in density forecast performance after the crisis. At short forecast horizons, the difference in log scores for GDP growth density forecasts that implicitly or explicitly take into account a broad information set – Survey-augmented COMPASS, the Statistical Suite and the *Inflation Report* – and those that do not – the two other variants of COMPASS – is larger in the post-crisis period. This observation suggests that conditioning on timely indicators that may summarise a broad range of information, such as survey growth expectations, or a broad information set, can improve forecast accuracy especially in times of heightened volatility.[41]

It is also informative to investigate how the logarithmic scores have changed over time. Figure 14 calculates the logarithmic scores over a rolling window of four years for GDP growth and inflation density forecasts at a horizon 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 the accuracy of the density forecasts. There is a much less clear pattern for the rolling log scores computed from the inflation density forecasts. The scores of the inflation density forecasts for both Survey-augmented COMPASS and the *Inflation Report* decreased sharply during the crisis, while the log scores for the other three methods did not.

Figure 15 reports the fluctuation test statistic: the difference between average MSFE calculated over a rolling window of 28 quarters and standardised by an estimate of the long-run variance. The reported test statistics are relative to the Survey-augmented COMPASS, such that a

[40] See Appendix B for empirical size and power of the Diebold-Mariano test in small samples.

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

negative value of the test statistic means that this model performs better in terms of average MSFE. Casual inspection of Figure 15 indicates that relative forecasting performance has varied over the sample period for the inflation forecasts: the relative performance of the Statistical Suite and the *Inflation Report* declines over the sample, while the relative performance of other COMPASS variants improves during the middle part of the sample. For GDP growth, the relative performance of these other variants is approximately constant, while the relative performance of the Statistical Suite and the *Inflation Report* declines. Note that none of these fluctuation test statistics is statistically significant, which may in part reflect the relatively small sample size.

5.5 Interpreting the forecast performance of COMPASS through the lens of forecast error decompositions

We noted above that all variants of COMPASS underestimate inflation and overestimate GDP growth in the post-crisis period. This section uses forecast error decompositions to explore why this is the case. In particular, we take advantage of the structure of COMPASS to decompose the forecast errors observed over the crisis period into identified shocks.[42] Technically, this is computed as the difference between a shock-based decomposition of the data (including data over the forecast horizon) and a shock-based decomposition of the forecast under consideration (augmented with data prior to the forecast origin) with the shocks in both decompositions computed using the Kalman filter and smoother.[43] Using this method, we decompose the forecast errors for inflation and GDP growth for a forecast originating in 2009Q1 using Survey-augmented COMPASS.[44] We group the shocks into eight categories as explained in Appendix A.

Figure 16a reports the shock-based decomposition of the errors in the 2009Q1 forecast for GDP growth. The largest contributions to the negative forecast errors over the first six quarters are from productivity shocks (and a permanent productivity shock in particular) with a smaller role for a set of demand shocks (including the domestic risk premium shock and a world demand shock).

The role of productivity shocks is not surprising here given the unusually persistent weakness in measured labour productivity observed throughout the post-crisis period (see Barnett et al. (2014)). Within COMPASS, the permanent productivity shock is the only one that reduces the trend level of GDP permanently, and by extension the levels of other expenditure components.[45] As such, the mechanism plays an important part in accounting for the fall in GDP that was recorded in the data; in its absence, the model forecasts would have been much less accurate over this period.

[42] In related work, Hackworth et al. (2013) analyse the possible source of MPC forecast errors in the post-crisis period.

[43] Given a time series of the smoothed shocks, it is straightforward to compute a shock-based decomposition by summing the historical effect of each shock at each point in time (and adding in the (diminishing) effect of an initial condition for the state). See Section 6.2.5 of Burgess et al. (2013).

[44] For comparison, we also report forecast error decompositions for 2003Q3 (Figures 16c and 16d).

[45] This reflects an assumption that they follow difference-stationary processes and are cointegrated.

Although COMPASS does not include financial frictions, making it difficult to extract the contribution of credit and financial shocks over this period, the constellation of shocks identified here appears to be consistent with a financial-type shock.[46] In particular, credit supply shocks are likely to affect both the demand and supply side of the economy. For example, tighter credit conditions may impede the efficiency with which resources are allocated in the economy, thereby reducing its supply capacity. On the demand side, tighter credit conditions also affect the cost of credit for households, as captured by the domestic risk premium shock. Finally, given the global nature of the financial crisis, it is not surprising that world demand shocks should play an important role in explaining GDP growth forecast errors.

The forecast errors on the inflation point forecast made in 2009Q1 are explained through the lens of $COMPASS^{GDP^e}$ by monetary policy shocks pushing up on inflation, on the one hand, and domestic markup shocks pulling down, on the other. As documented by Burgess et al. (2013), energy price shocks[47] are likely to be captured by, among other things, domestic markup shocks. The presence of markup shocks in explaining the forecast error over this period is therefore consistent with the substantial fall in energy prices that occurred during the second half of 2008. On the other side, $COMPASS^{GDP^e}$ explains the bulk of the unforecast strength in inflation over this period as a monetary policy shock, which could be consistent with the exceptionally loose stance of policy over this period.[48]

6 Conditioning DSGE model forecasts

As discussed in Section 2.3, the *Inflation Report* forecasts are conditioned on some assumptions about the future paths of a set of variables. In this section, we evaluate how the application of this conditioning information affects the accuracy of forecasts from COMPASS. We investigate two different sets of conditioning paths:

1. CP^{BOE}: The policy interest rate is constrained to follow a particular path derived from market expectations; the exchange rate is constrained to follow a path that is an average of a forecast from a random walk and a static UIP condition;[49] government spending is assumed to evolve in line with the government's latest announced fiscal plans; and world demand and export prices are constrained to match forecasts produced by an international forecast team on behalf of the MPC.

[46] As explained by Burgess et al. (2013), shock decompositions can only be interpreted as truly structural if the model on which they depend is correctly specified. In practice, all models are misspecified and the shocks cannot be interpreted as truly structural. In fact, the shocks are likely to be correlated and shocks that are not included in the model are likely to be captured by a correlated sequence of shocks that are included.

[47] That is, unexplained variation in energy prices that is not attributable to shocks that induce a change in demand for energy.

[48] The effects of quantitative easing are approximated in a shadow measure of the interest rate – see Section 2.1 for a discussion.

[49] With domestic and foreign interest rates following a market path.

2. CP^{π}: The inflation nowcast and the inflation forecast one quarter ahead are constrained to match a short-term forecast produced by Bank staff.

For the purposes of these experiments, we use the Plain version of COMPASS and apply all conditioning information using unanticipated shocks under two different schemes:[50]

(i) we use the same pre-selected set of shocks used by convention in constructing the MPC's judgemental forecasts using COMPASS as part of the forecast process at the Bank;[51]

(ii) we use the full range of model shocks to impose the paths, solving the identification problem that this poses, by selecting the minimal variance combination ($CP^{BOEfull}$, $CP^{\pi full}$).[52]

Below, we document material differences in the accuracy of the forecasts that arise from these two alternative schemes, compared to the baseline case of using the Plain version of COMPASS without any conditioning information.[53] Figures 17a and 17b compare forecast accuracy when we vary the conditioning information applied, and the method of applying the information.

Imposing the standard conditioning information with pre-specified shocks (CP^{BOE}) reduces forecast accuracy at almost all horizons. These differences are statistically significant for GDP growth at horizons of one and two years, at 5% and 10% significance respectively. By contrast, applying the same conditioning information but using *all* of the shocks ($CP^{BOEfull}$) improves the accuracy of GDP growth forecasts up to two years ahead (statistically significant at one year at a 10% level), and inflation forecasts at almost all horizons.

This difference reflects the choice of shocks used and the corresponding economic interpretation of the conditioning information. For example, the interpretation of conditioning the forecast on market expectations of the policy rate using the pre-specified set of shocks is that market participants have a similar view of the state of the economy and the outlook, but a different view of how monetary policy will evolve.[54] By contrast, imposing the conditioning information using all shocks affords a role for market participants having a different view of the outlook for

[50] In doing so, we use the toolkit described in Section 6 of Burgess et al. (2013)

[51] CP^{BOE} are imposed using a monetary policy shock, exchange rate risk premium shock, government spending shock, world demand shock and world export price shock. CP^{π} is imposed using a final output markup shock.

[52] See Burgess et al. (2013), Appendix C for a derivation and discussion.

[53] 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 $h-1$ which is often recommended in the forecast evaluation literature (Diebold and Mariano (1995)).There are broadly two methods to compute density forecasts conditional on future paths of a set of variables: Banbura, Giannone, and Lenza (2014) develop an algorithm based on a Kalman filter with missing observations. In contrast to the methodology in Waggoner and Zha (1999), the Kalman filter computes conditional forecasts recursively which makes it applicable to large models. An alternative approach is to project the forecast densities in the space of distributions that is consistent with the conditioning path (Giacomini and Ragusa (2011)). In future versions of this paper, we intend to update our analysis to include conditional density forecast evaluation.

[54] The forecast is also conditioned on a path for the exchange rate, government spending and the outlook for the world. All of these paths are imposed at the same time, so it is the entire constellation of shocks being used – see footnote 51 – that affects the interpretation of the information in all of the conditioning paths. However, in practice, the monetary policy shock tends to explain the bulk of the interest rate conditioning path, so this description is approximately true.

the economy, as well as monetary policy. The post-crisis period illustrates why this matters. Throughout this period, market expectations for the policy rate are lower than forecasts from the Plain COMPASS. When applying CP^{BOE}, we introduce negative monetary policy shocks which have the effect of boosting the model's forecasts for GDP growth and inflation. This worsens the model's accuracy in forecasting GDP growth. But applying a broader set of shocks causes the model to attribute the lower policy path to shocks which *reduce* the model's forecasts for GDP growth and inflation, thereby delivering an endogenously lower interest rate forecast.

Conditioning on the Bank staff's inflation nowcasts and short-term forecasts improves the accuracy of the model's forecast for inflation up to one year ahead, regardless of whether the standard shock is used to impose the conditioning information or whether all the shocks are used. Neither conditioning assumption has much effect on the accuracy of the model's GDP growth forecasts. As we noted earlier, the *Inflation Report* forecasts are conditioned on these staff nowcasts.

When investigating the time-variation in conditional RMSFE at a forecast horizon of one year, we find that imposing CP^{BOE} worsens inflation and GDP growth forecast accuracy relative to almost all the other models at almost all points in time. By contrast, $CP^{BOEfull}$ has the lowest RMSFE for GDP growth during the crisis and thereafter. This suggests that the interpretation given to the yield curve when all of the shocks are used is particularly appealing during the crisis. For inflation, there is a much less clear pattern, though $CP^{BOEfull}$ has the lowest RMSFE for much of the sample.

7 Conclusions

This paper has evaluated the accuracy of real-time forecasts for inflation and GDP growth from an estimated DSGE model for the United Kingdom before, during and after the financial crisis, compared to a statistical benchmark, and the MPC's forecasts in the *Inflation Report*. At shorter horizons, the MPC's judgemental forecasts – which implicitly take account of a broad information set – are more accurate. At longer horizons, the model-based forecasts become more accurate, with forecasts for GDP growth from the Statistical Suite and forecasts for inflation from the DSGE model the most accurate respectively.

The accuracy of all forecasts fell during the financial crisis. Performance at the peak of the crisis was poor, and even in the years that followed, the *Inflation Report* and several model-based methods tended to over-predict GDP growth and under-predict inflation. This deterioration was particularly marked for the GDP growth forecasts from the DSGE model. In line with other results in the literature, the accuracy of the DSGE model's forecasts for GDP is significantly improved if its information set is augmented to include more timely, forward-looking information, in the form of survey expectations of GDP growth. The performance of GDP forecasts from this alternative model is on a par with those of the Statistical Suite and *Inflation Report*. This is particularly important in improving the accuracy of the model's forecasts during the financial crisis, when

timely, forward-looking information is likely to have been particularly valuable.

More generally, the recent financial crisis has posed new challenges for DSGE modelling and forecasting. From a modelling perspective, the current generation of DSGE models is not well-suited to capturing the implications of large financial shocks that may have non-linear effects, particularly when policy rates become constrained by the zero lower bound. From a forecasting perspective, dealing with possible structural breaks in point and density forecasting is challenging, especially in real time. With research underway that aims to address these challenges, it is probable that DSGE modelling and forecasting will undergo significant changes in the future.

References

AIOLFI, M., C. CAPISTRAN, AND A. TIMMERMANN (2010): "Forecast Combinations," CREATES Research Paper 2010-21, School of Economics and Management, University of Aarhus.

ALESSANDRI, P. AND H. MUMTAZ (2013): "Financial Conditions and Density Forecasts for US Output and Inflation," *Centre for Central Banking Studies Joint Research Paper No. 4*.

AMISANO, G. AND R. GIACOMINI (2007): "Comparing Density Forecasts via Weighted Likelihood Ratio Tests," *Journal of Business & Economic Statistics*, 25, 177–190.

BANBURA, M., D. GIANNONE, AND M. LENZA (2014): "Conditional Forecasts and Scenario Analysis with Vector Autoregressions for Large Cross-Sections," *ECARES working paper*.

BANK OF ENGLAND (2000): *Economic models at the Bank of England: September 2000 update*, London: Bank of England.

BARNETT, A., S. BATTEN, A. CHIU, J. FRANKLIN, AND M. SEBASTIÁ-BARRIEL (2014): "The UK productivity puzzle," *Bank of England Quarterly Bulletin*, 54, 114–119.

BERKOWITZ, J. (2001): "Testing density forecasts, with applications to risk management," *Journal of Business & Economic Statistics*, 19, 465–474.

BERNANKE, B. S., M. GERTLER, AND S. GILCHRIST (1999): "The financial accelerator in a quantitative business cycle framework," in *Handbook of Macroeconomics*, ed. by J. B. Taylor and M. Woodford, Elsevier, vol. 1 of *Handbook of Macroeconomics*, chap. 21, 1341–1393.

BURGESS, S., E. FERNANDEZ-CORUGEDO, C. GROTH, R. HARRISON, F. MONTI, K. THEODORIDIS, AND M. WALDRON (2013): "The Bank of England's Forecasting Platform: COMPASS, MAPS, EASE and the Suite of Models," *Bank of England working paper*.

CHRISTIANO, L. J., M. EICHENBAUM, AND C. L. EVANS (2005): "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy," *Journal of Political Economy*, 113, 1–45.

CHRISTOFFEL, K., G. COENEN, AND A. WARNE (2010): "Forecasting with DSGE Models," *ECB working paper*.

CLARK, T. AND M. MCCRACKEN (2013): "Advances in Forecast Evaluation," in *Handbook of Economic Forecasting*, ed. by G. Elliott and A. Timmermann, Elsevier, vol. 2, Part B, 1107 – 1201.

CLARK, T. E. AND M. W. MCCRACKEN (2014): "Evaluating Conditional Forecasts from Vector Autoregressions," *SSRN working paper*.

CLEMENTS, M. AND D. HENDRY (1998): *Forecasting Economic Time Series*, Cambridge, UK: Cambridge University Press.

COMIN, D. AND M. GERTLER (2006): "Medium-Term Business Cycles," *American Economic Review*, 96, 523–551.

CORRADI, V. AND N. R. SWANSON (2006): "Predictive Density Evaluation," Handbook of economic forecasting, Elsevier.

DEL NEGRO, M. AND F. SCHORFHEIDE (2012): "DSGE Model-based Forecasting," *SSRN working paper*.

DIEBOLD, F. X. AND R. S. MARIANO (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–63.

EDGE, R. AND R. GÜRKAYNAK (2011): "How Useful are Estimated DSGE Model Forecasts?" *Available at SSRN 1810075*.

EDGE, R. M., M. T. KILEY, AND J.-P. LAFORTE (2010): "A Comparison of Forecast Performance between Federal Reserve Staff Forecasts, Simple Reduced-form Models, and a DSGE Model," *Journal of Applied Econometrics*, 25, 720–754.

FERRONI, F. (2011): "Trend Agnostic One-Step Estimation of DSGE Models," *The BE Journal of Macroeconomics*, 11, 1–36.

FILIPPO, F. (2011): "Trend Agnostic One-Step Estimation of DSGE Models," *The B.E. Journal of Macroeconomics*, 11, 1–36.

GIACOMINI, R. AND G. RAGUSA (2011): "Incorporating theoretical restrictions into forecasting by projection methods," *CEPR discussion paper*.

GIACOMINI, R. AND B. ROSSI (2010): "Forecast Comparisons in Unstable Environments," *Journal of Applied Econometrics*, 25, 595–620.

GIACOMINI, R. AND H. WHITE (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578.

GÜRKAYNAK, R., B. KISACIKOGLU, AND B. ROSSI (2013): "Do DSGE Models Forecast More Accurately Out-of-Sample than VAR Models?" *mimeo*.

HAAN, W. J. D. AND A. T. LEVIN (1996): "A Practitioner's Guide to Robust Covariance Matrix Estimation," *NBER working paper*.

HACKWORTH, C., A. RADIA, AND N. ROBERTS (2013): "Understanding the MPC's forecast performance since mid-2010," *Bank of England Quarterly Bulletin*, 53, 336–350.

HARRISON, R., K. NIKOLOV, M. QUINN, A. SCOTT, AND R. THOMAS (2005): *The Bank of England Quarterly Model*, London: The Bank of England.

IVERSEN, J., S. LASEEN, H. LUNDVALL, AND U. SÖDERSTRÖM (2013): "Monetary Policy Modelling in Times of Financial Turmoil: The Case of the Sveriges Riksbank," *mimeo*.

JOYCE, M., M. TONG, AND R. WOODS (2011): "The United Kingdom's quantitative easing policy: design, operation and impact," *Bank of England Quarterly Bulletin*, 51, 200–212.

KAPETANIOS, G., V. LABHARD, AND S. PRICE (2008): "Forecast combination and the Bank of England's suite of statistical forecasting models," *Economic Modelling*, 25, 772–792.

NEWEY, W. K. AND K. D. WEST (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703.

OULTON, N. AND M. SEBASTIÁ-BARRIEL (2013): "Long and short-term effects of the financial crisis on labour productivity, capital and output," *Bank of England Working Paper*, 470.

PETROVA, K. (2013): "Real Time Forecasting in the Recent Crisis with a Medium-Sized DSGE Model," *mimeo*.

SMETS, F. AND R. WOUTERS (2003): "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area," *Journal of the European Economic Association*, 1, 1123–1175.

WAGGONER, D. F. AND T. ZHA (1999): "Conditional Forecasts In Dynamic Multivariate Models," *The Review of Economics and Statistics*, 81, 639–651.

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.

Table 1: Diebold-Mariano test of equal relative forecasting ability for inflation

a) 1 year ahead forecasts

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	0.467	0.265	-0.768	0.368
$COMPASS^{TVT}$	.	.	-0.989	-0.949	-0.165
$COMPASS^{PLAIN}$	.	.	.	-0.806	0.135
IR	.	.	.	.	0.939
Stat. Suite	.	.	.	.	.

b) 2 years ahead forecast

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	0.073	-0.067	-0.999	-1.276
$COMPASS^{TVT}$	.	.	-0.750	-0.874	-1.002
$COMPASS^{PLAIN}$	.	.	.	-0.844	-0.924
IR	.	.	.	.	0.655
Stat. Suite	.	.	.	.	.

Table 2: Diebold-Mariano test of equal relative forecasting ability for GDP growth

a) 1 year ahead forecasts

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	-1.986**	-1.895*	0.354	-1.347
$COMPASS^{TVT}$	.	.	-0.623	1.725*	2.279**
$COMPASS^{PLAIN}$	.	.	.	1.815*	2.128**
IR	.	.	.	.	-0.792
Stat. Suite	.	.	.	.	.

b) 2 years ahead forecast

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	-0.734	-1.936*	-0.397	1.810*
$COMPASS^{TVT}$	.	.	-0.048	0.087	1.340
$COMPASS^{PLAIN}$	.	.	.	0.193	2.235**
IR	.	.	.	.	0.778
Stat. Suite	.	.	.	.	.

Notes: Test statistics are shown. 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 5% (10%). A value 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 3: Giacomini-White test of equal conditional relative forecasting ability at a horizon of 1 year (oil price)

a) Inflation

i) Test statistic

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	0.855	0.747	2.094	1.279
$COMPASS^{TVT}$	.	.	2.455	1.727	3.095
$COMPASS^{PLAIN}$	.	.	.	1.273	2.813
IR	.	.	.	.	2.085
Stat. Suite	.	.	.	.	.

ii) Range of the oil price where the column model is more accurate

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$		[Min,8.16]	[Min,6.65]	[Min,-8.91]	[Min,6.48]
$COMPASS^{TVT}$			[ ]	[ ]	[Min,2.59]
$COMPASS^{PLAIN}$				[ ]	[Min,6.14]
IR					[Min,23.34]
Stat. Suite					

b) GDP growth

i) Test statistic

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	9.687^{--}	8.392^{--}	0.442	1.733
$COMPASS^{TVT}$	.	.	0.390	6.453^{++}	16.557^{++}
$COMPASS^{PLAIN}$	.	.	.	7.019^{++}	17.379^{++}
IR	.	.	.	.	1.227
Stat. Suite	.	.	.	.	.

ii) Range of the oil price where the column model is more accurate

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$		[ ]	[ ]	[-19.14,Max]	[ ]
$COMPASS^{TVT}$			[ ]	[Min,Max]	[Min,Max]
$COMPASS^{PLAIN}$				[Min,Max]	[Min,Max]
IR					[ ]
Stat. Suite					

Notes: Test statistics are shown. 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) 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 10% (5%).

Table 4: Giacomini-White test of equal conditional relative forecasting ability at a horizon of 1 year (corporate credit spread)

a) Inflation

i) Test statistic

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	1.257	2.252	1.407	2.144
$COMPASS^{TVT}$	.	.	1.017	1.539	0.027
$COMPASS^{PLAIN}$	.	.	.	1.629	0.027
IR	.	.	.	.	2.043
Stat. Suite	.	.	.	.	.

ii) Range of the corporate credit spread where the column model is more accurate

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$		[1.58,Max]	[1.69,Max]	[ ]	[1.63,Max]
$COMPASS^{TVT}$			[5.0929,Max]	[Min,1.2824]	[]
$COMPASS^{PLAIN}$				[Min,1.35]	[Min,3.41]
IR					[1.31,Max]
Stat. Suite					

b) GDP growth

i) Test statistic

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	4.034	4.238	0.169	2.543
$COMPASS^{TVT}$	.	.	4.148	3.412	8.052^{++}
$COMPASS^{PLAIN}$	.	.	.	3.733	7.838^{++}
IR	.	.	.	.	0.933
Stat. Suite	.	.	.	.	.

ii) Range of the corporate credit spread where the column model is more accurate

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$		[ ]	[ ]	[Min,Max]	[Min,0.83]
$COMPASS^{TVT}$			[Min,1.34]	[Min,Max]	[Min,4.51]
$COMPASS^{PLAIN}$				[Min,Max]	[Min,Max]
IR					[ ]
Stat. Suite					

Notes: Test statistics are shown. The test functions are $(1, X_t)$ where X_t is the corporate credit spread. A Newey-West estimator with an optimally chosen bandwidth is used to estimate the long-run variance.A plus (minus) 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 10% (5%).

Table 5: Amisano-Giacomini test of equal density forecasts for inflation

a) 1 year ahead forecasts

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	-1.372	-1.036	0.217	-2.036**
$COMPASS^{TVT}$	.	.	1.547	0.880	-1.795*
$COMPASS^{PLAIN}$	.	.	.	0.680	-1.761*
IR	.	.	.	.	-1.806*
Stat. Suite	.	.	.	.	.

b) 2 years ahead forecast

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	1.606	1.573	1.296	-1.555
$COMPASS^{TVT}$	.	.	1.753*	0.985	-1.703*
$COMPASS^{PLAIN}$	.	.	.	0.869	-1.603
IR	.	.	.	.	-1.665*
Stat. Suite	.	.	.	.	.

Table 6: Amisano-Giacomini test of equal density forecasts for GDP growth

a) 1 year ahead forecasts

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	2.085**	1.984**	1.462	0.969
$COMPASS^{TVT}$	.	.	0.776	-0.725	0.297
$COMPASS^{PLAIN}$	.	.	.	-1.556	0.055
IR	.	.	.	.	0.642
Stat. Suite	.	.	.	.	.

b) 2 years ahead forecast

	$COMPASS^{GDP^e}$	$COMPASS^{TVT}$	$COMPASS^{PLAIN}$	IR	Stat. Suite
$COMPASS^{GDP^e}$	.	0.014	2.519**	1.180	0.846
$COMPASS^{TVT}$	.	.	0.516	1.070	0.803
$COMPASS^{PLAIN}$	.	.	.	1.070	0.765
IR	.	.	.	.	0.303
Stat. Suite	.	.	.	.	.

Notes: Test statistics are shown. 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 5% (10%). A value smaller (larger) than zero indicates that the model in the corresponding column generates more (less) accurate forecasts than that listed in the corresponding row.

Table 7: Berkowitz test for forecast rationality

a) 1 year ahead forecasts

	Inflation	GDP growth
$COMPASS^{GDP^e}$	38.902	86.742
$COMPASS^{TVT}$	43.103	159.075
$COMPASS^{PLAIN}$	49.151	230.228
IR	22.156	45.580
Stat. Suite	25.471	55.029

b) 2 years ahead forecasts

	Inflation	GDP growth
$COMPASS^{GDP^e}$	103.045	126.935
$COMPASS^{TVT}$	92.455	138.372
$COMPASS^{PLAIN}$	90.434	137.755
IR	48.417	55.014
Stat. Suite	72.553	79.570

Notes: Test statistics are reported. The asymptotic distribution is $\chi^2(2)$ and all test statistics are significant at 1 percent.

Table 8: Diebold-Mariano test of equal relative forecasting ability for inflation for conditional forecasts

a) 1 year ahead forecasts

	$COMPASS^{PLAIN}$	CP^{BOE}	$CP^{BOEfull}$	CP^{π}	$CP^{\pi full}$
$COMPASS^{PLAIN}$	.	-1.099	0.777	-0.378	0.254
CP^{BOE}	.	.	1.116	1.111	1.228
$CP^{BOEfull}$	.	.	.	-1.086	-0.622
CP^{π}	.	.	.	.	0.499
$CP^{\pi full}$	.	.	.	.	.

b) 2 years ahead forecast

	$COMPASS^{PLAIN}$	CP^{BOE}	$CP^{BOEfull}$	CP^{π}	$CP^{\pi full}$
$COMPASS^{PLAIN}$	.	-1.264	0.911	-1.982	1.143
CP^{BOE}	.	.	1.250	1.220	1.273
$CP^{BOEfull}$	.	.	.	-1.024	-0.910
CP^{π}	.	.	.	.	1.775
$CP^{\pi full}$	.	.	.	.	.

Table 9: Diebold-Mariano test of equal relative forecasting ability for GDP growth

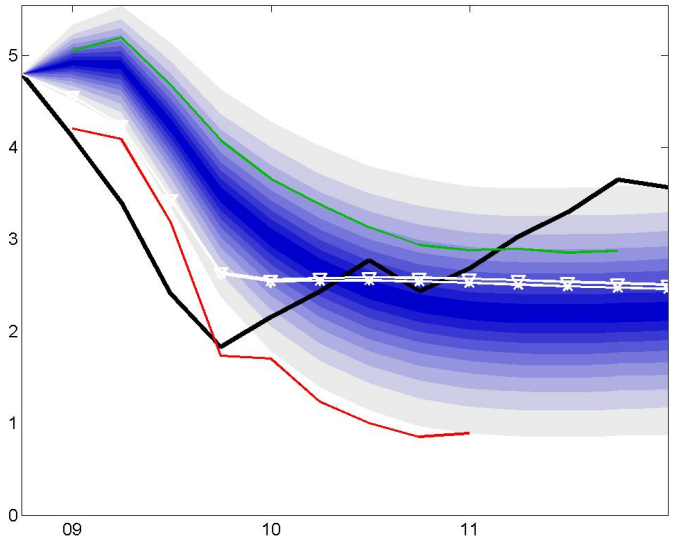
a) 1 year ahead forecasts

	$COMPASS^{PLAIN}$	CP^{BOE}	$CP^{BOEfull}$	CP^{π}	$CP^{\pi full}$
$COMPASS^{PLAIN}$	.	-2.084**	1.815*	-1.168	-0.089
CP^{BOE}	.	.	2.051**	2.081**	1.942*
$CP^{BOEfull}$	.	.	.	-2.057**	-2.658**
CP^{π}	.	.	.	.	0.111
$CP^{\pi full}$	.	.	.	.	.

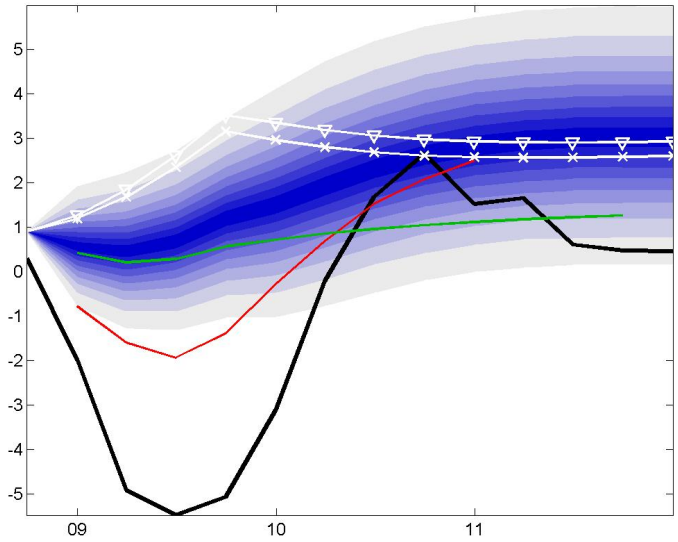
b) 2 years ahead forecast

	$COMPASS^{PLAIN}$	CP^{BOE}	$CP^{BOEfull}$	CP^{π}	$CP^{\pi full}$
$COMPASS^{PLAIN}$	.	-1.849*	0.799	1.191	-2.206**
CP^{BOE}	.	.	1.567	1.903*	1.883*
$CP^{BOEfull}$	.	.	.	-0.528	-2.029**
CP^{π}	.	.	.	.	-1.941*
$CP^{\pi full}$	.	.	.	.	.

Notes: Test statistics are shown. 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 5% (10%). A value 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

Figure 1: Inflation and GDP growth forecast in 2008Q4: Annual Output Growth ("first final" equivalent) in black, COMPASS$^{GDP^e}$ (COMPASS with survey growth expectations as observables) density forecast as blue fan, COMPASS^{PLAIN} (Burgess et al. (2013)) point forecasts in white with triangles, COMPASS^{TVT} (COMPASS with time-varying productivity trends) point forecasts in white with crosses, Statistical Suite forecasts in green and *Inflation Report* forecasts in red.

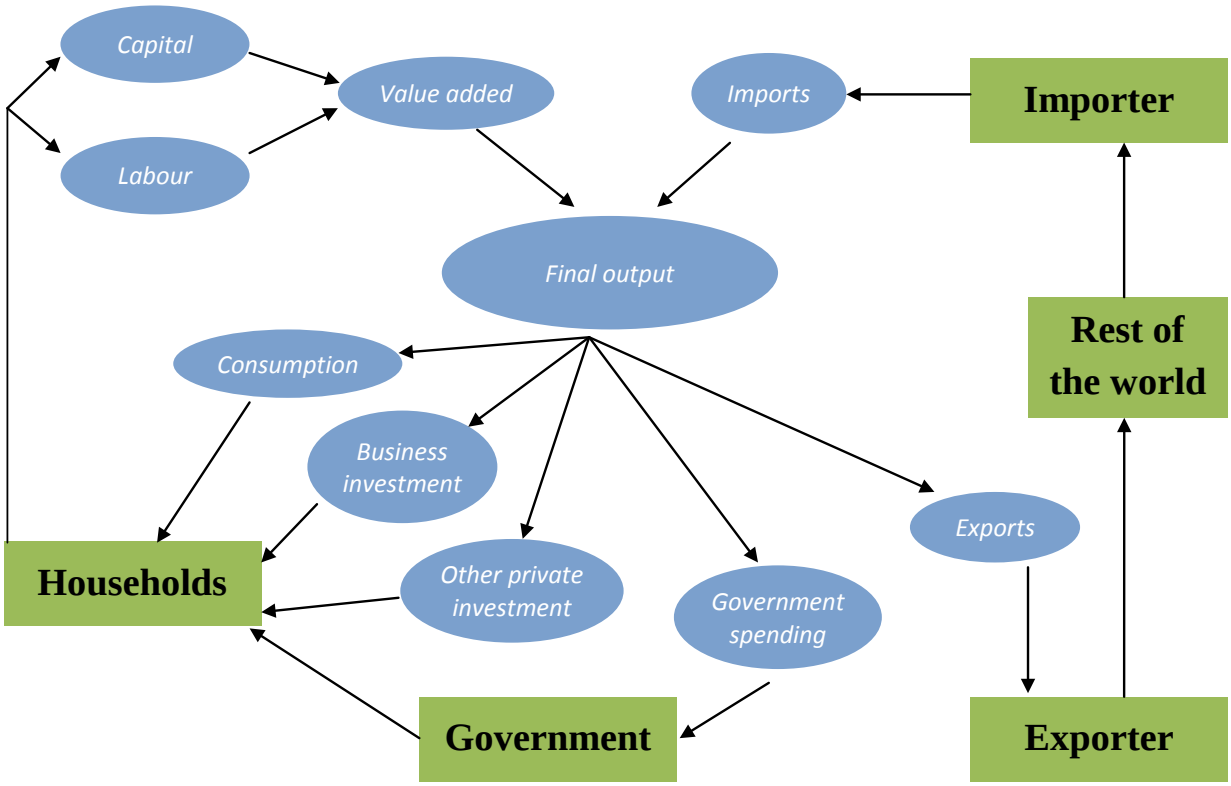


Figure 2: Overview of COMPASS. The *Central Organizing Model for Projection Analysis and Scenario Simulation* (COMPASS) is the main organizing framework for the *Inflation Report* forecasts at the Bank of England.

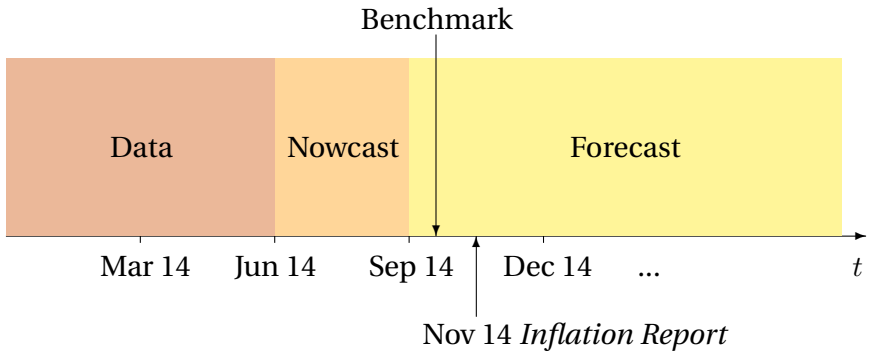
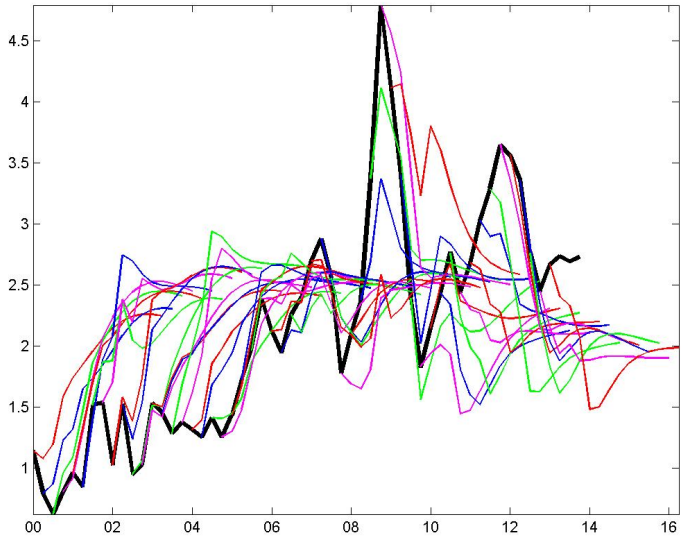
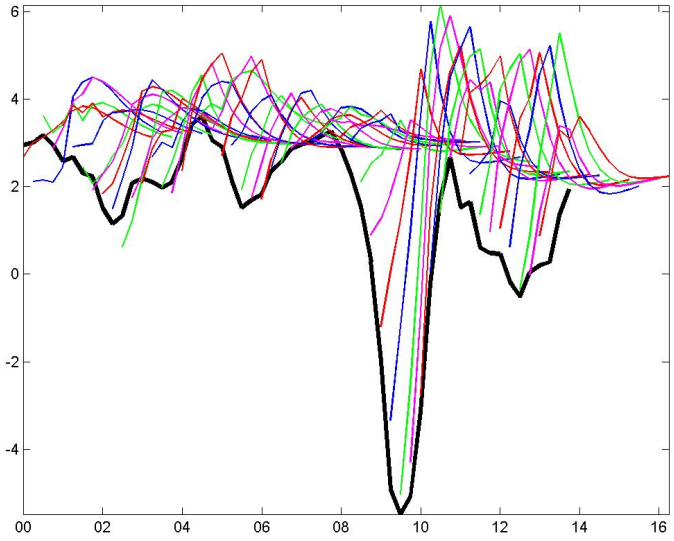


Figure 3: Time line of the forecast process at the Bank of England in case of the November 2014 forecast round

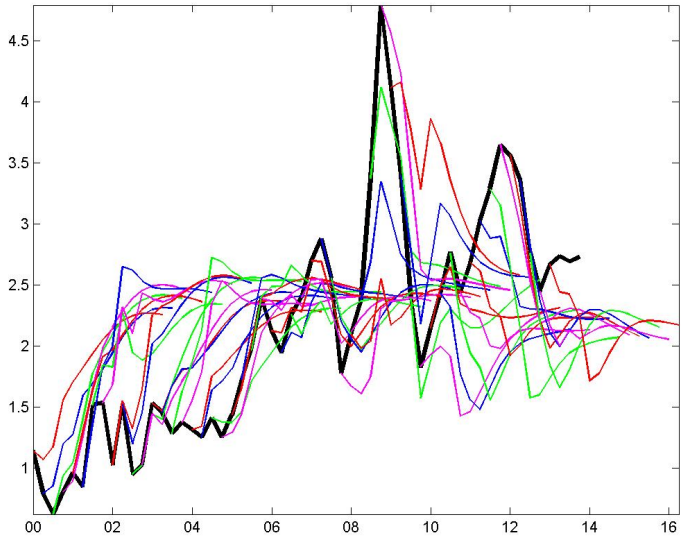


(a) Inflation

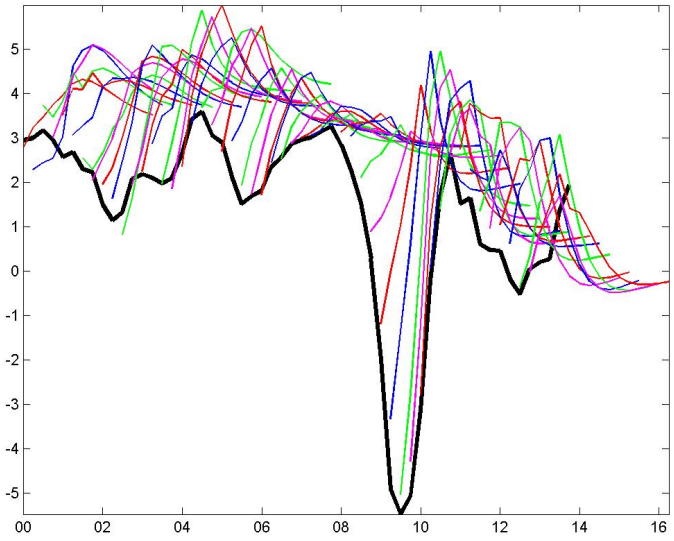


(b) GDP growth

Figure 4: Annual Inflation and GDP Growth ("first final" equivalent) in black and $COMPASS^{PLAIN}$ (estimated in real time) forecasts in color. November rounds in red, February rounds in blue, May rounds in green and July rounds in magenta.

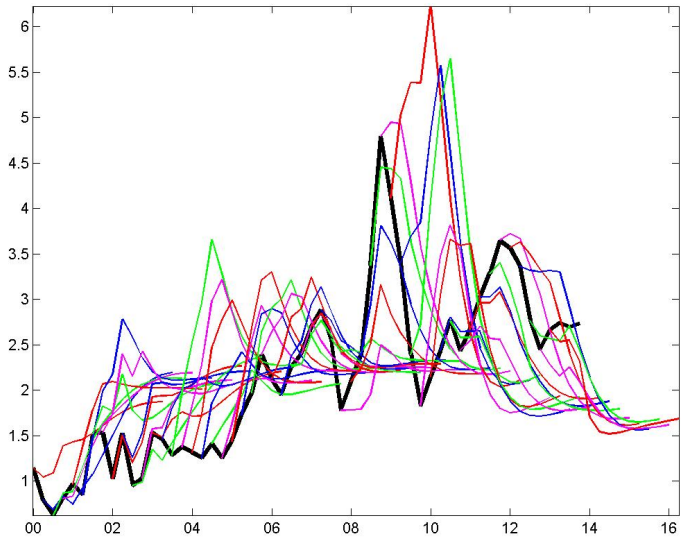


(a) Inflation

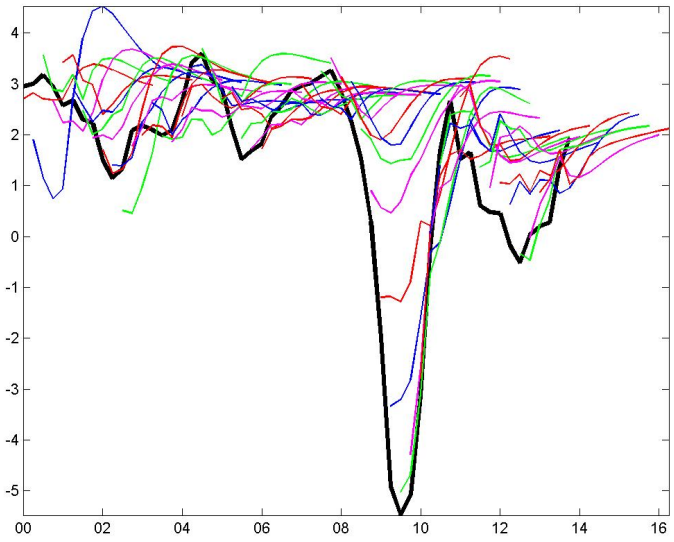


(b) GDP growth

Figure 5: Annual Inflation and GDP Growth ("first final" equivalent) in black and $COMPASS^{TVT}$ (estimated in real time) forecasts in color. November rounds in red, February rounds in blue, May rounds in green and July rounds in magenta.

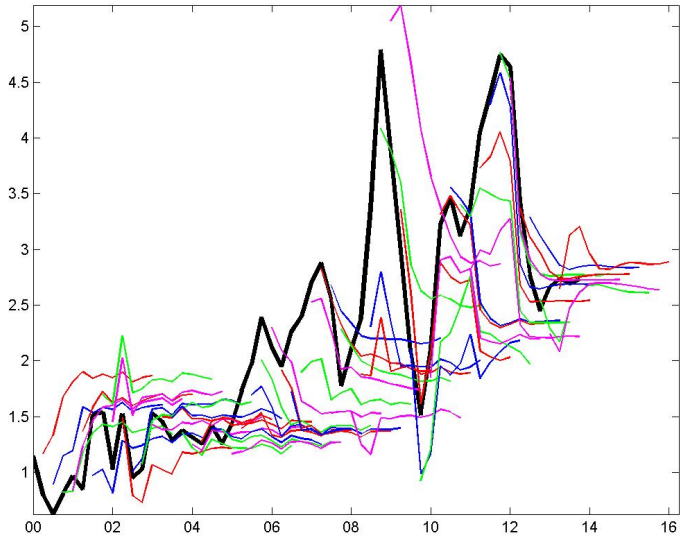


(a) Inflation

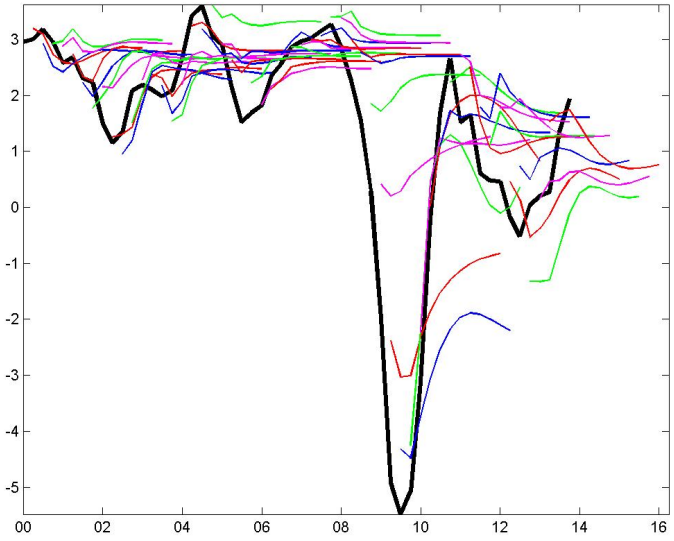


(b) GDP growth

Figure 6: Annual Inflation and GDP Growth ("first final" equivalent) in black and $COMPASS^{GDP^e}$ (estimated in real time) forecasts in color. November rounds in red, February rounds in blue, May rounds in green and July rounds in magenta.

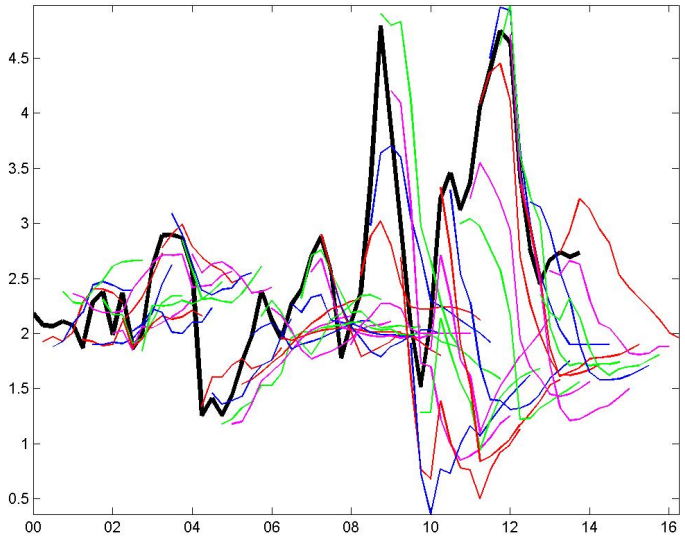


(a) Inflation

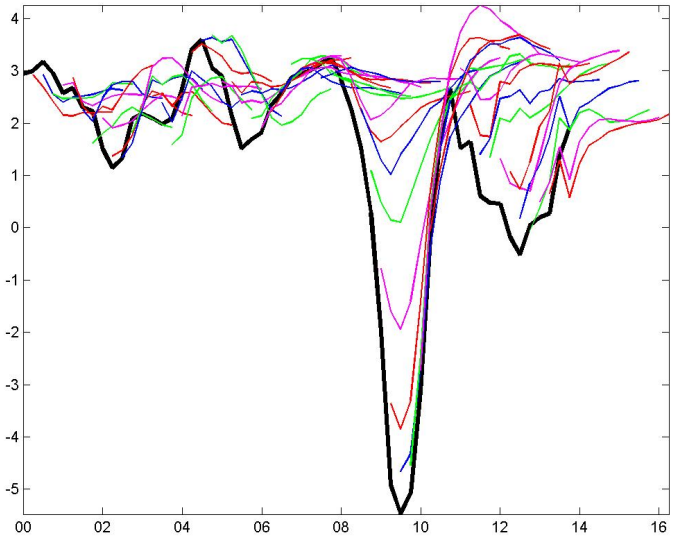


(b) GDP growth

Figure 7: Annual Inflation and GDP Growth ("first final" equivalent) in black and Statistical Suite (estimated in real time) forecasts in color. November rounds in red, February rounds in blue, May rounds in green and July rounds in magenta.

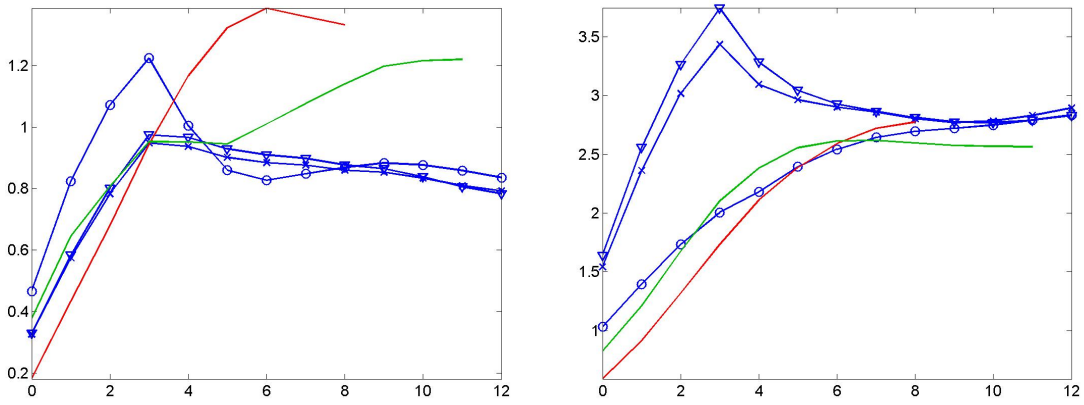


(a) Inflation



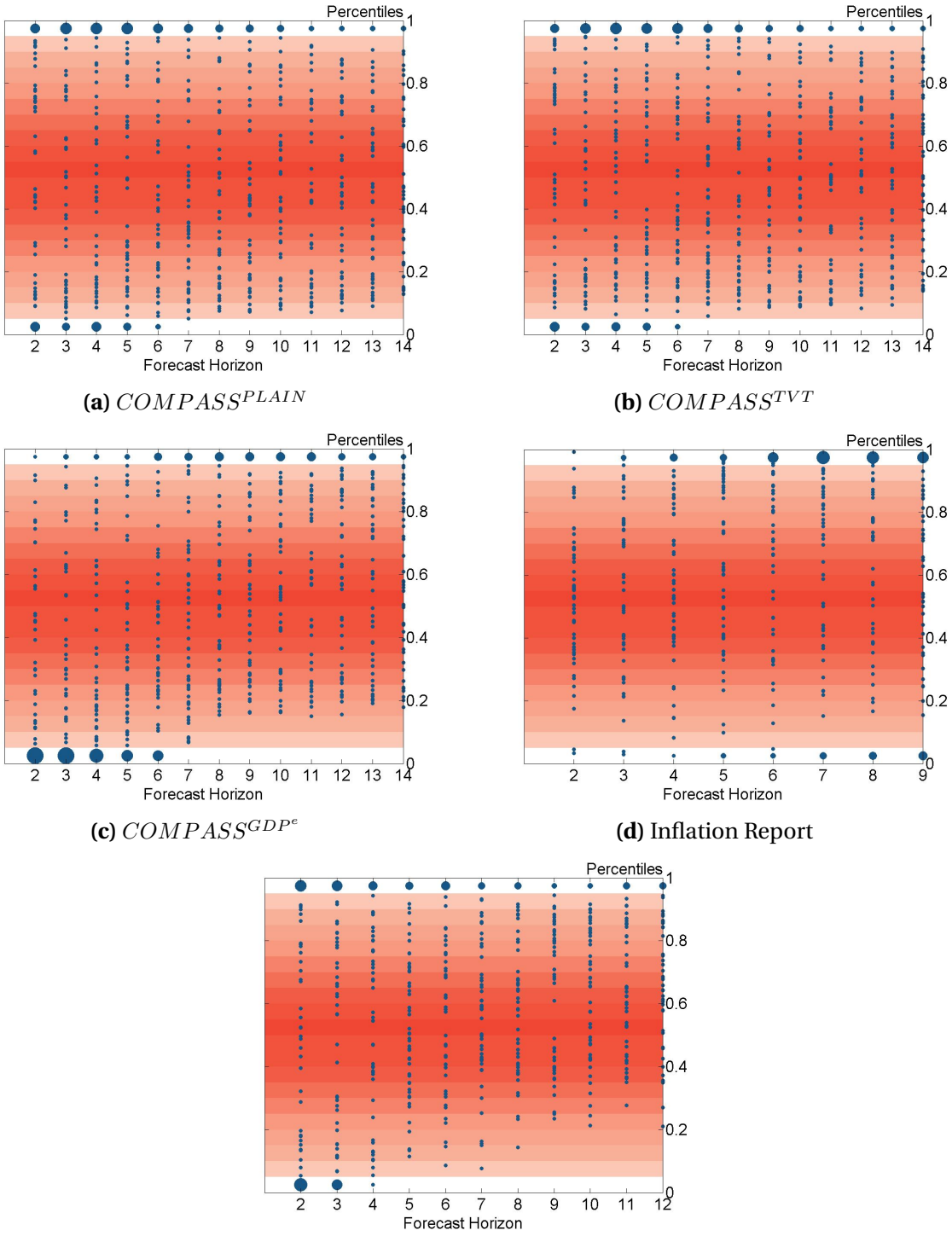
(b) GDP growth

Figure 8: Annual Inflation and GDP Growth ("first final" equivalent) in black and *Inflation Report* forecasts in color. November rounds in red, February rounds in blue, May rounds in green and July rounds in magenta.



(a) Inflation **(b)** GDP growth

Figure 9: Root mean squared forecast errors for the full sample period at different forecast horizons.COMPASS^{PLAIN} forecasts are in blue with triangles, COMPASS^{TVT} forecasts in blue with crosses, COMPASS$^{GDP^e}$ forecasts in blue with circles, Statistical Suite forecasts in green and *Inflation Report* forecasts in red.



(a) $COMPASS^{PLAIN}$

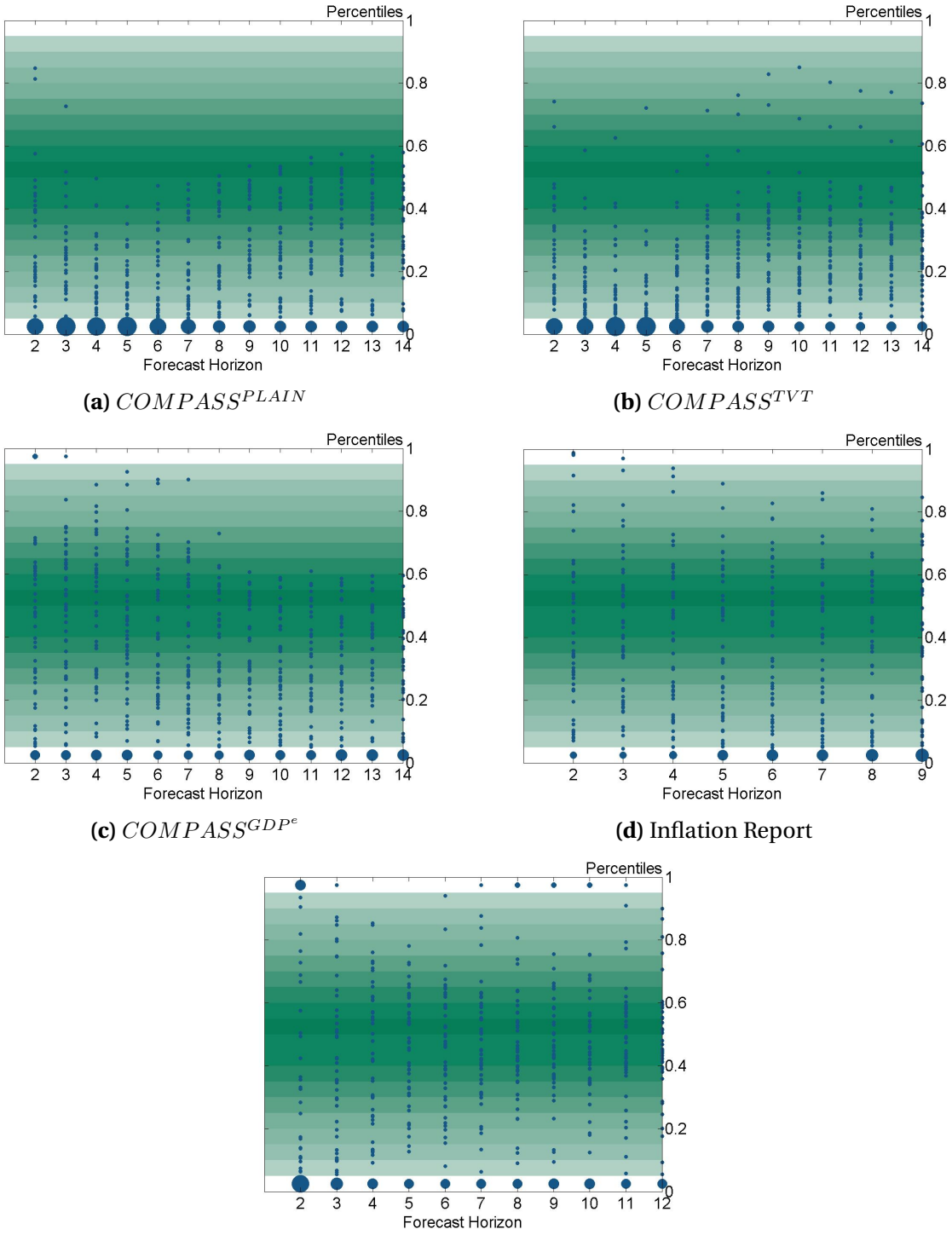
(b) $COMPASS^{TVT}$

(c) $COMPASS^{GDP^e}$

(d) Inflation Report

(e) Statistical Suite

Figure 10: Probability integral transforms for inflation.



(a) $COMPASS^{PLAIN}$

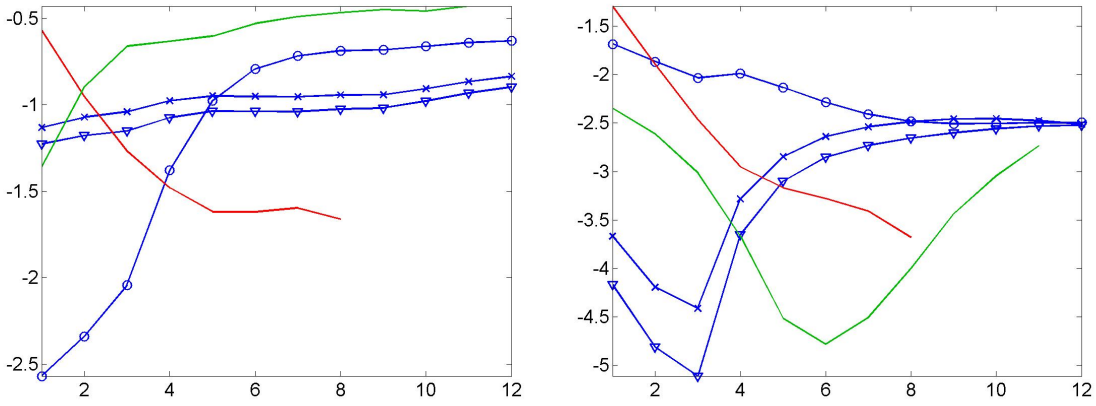
(b) $COMPASS^{TVT}$

(c) $COMPASS^{GDP^e}$

(d) Inflation Report

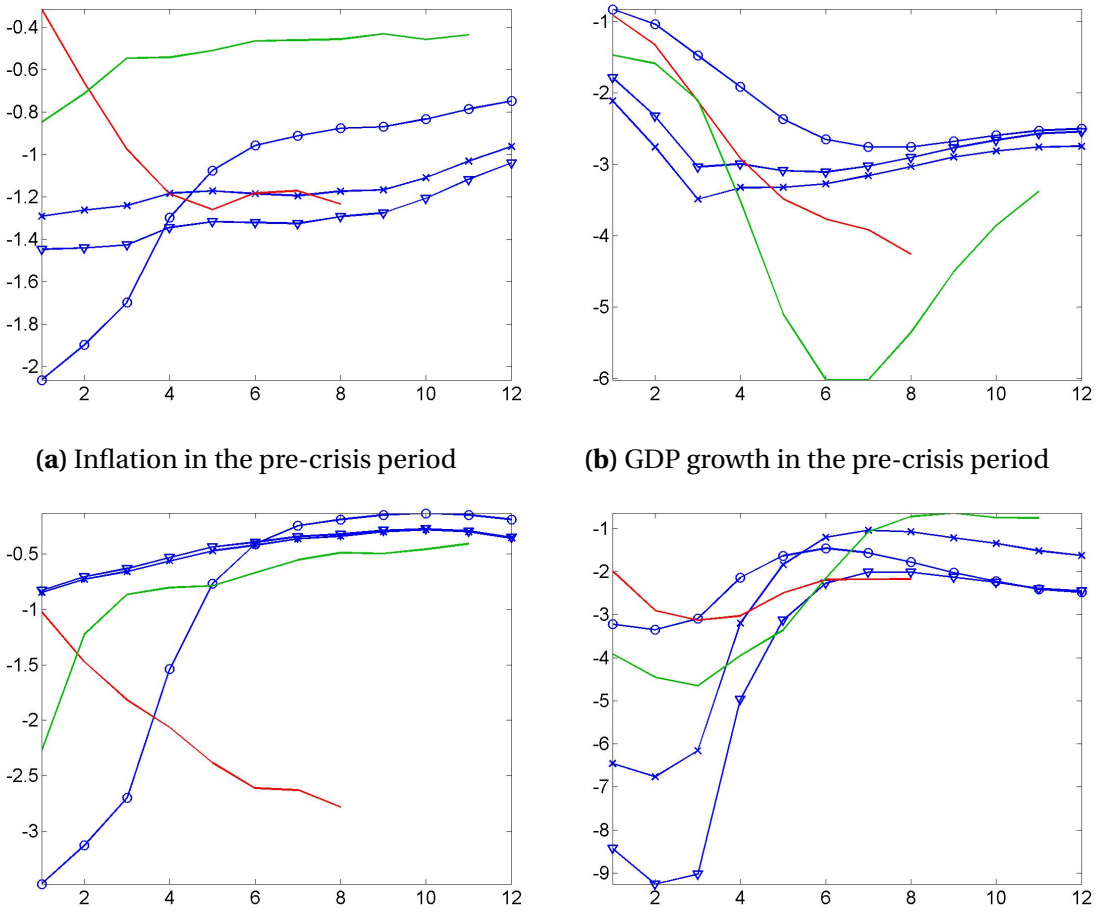
(e) Statistical Suite

Figure 11: Probability integral transforms for GDP growth.



(a) Inflation **(b)** GDP growth

Figure 12: Logarithmic scores at different forecast horizons. COMPASS^{PLAIN} forecasts are in blue with triangles, COMPASS^{TVT} forecasts in blue with crosses, COMPASS$^{GDP^e}$ forecasts in blue with circles, Statistical Suite forecasts in green and *Inflation Report* forecasts in red.



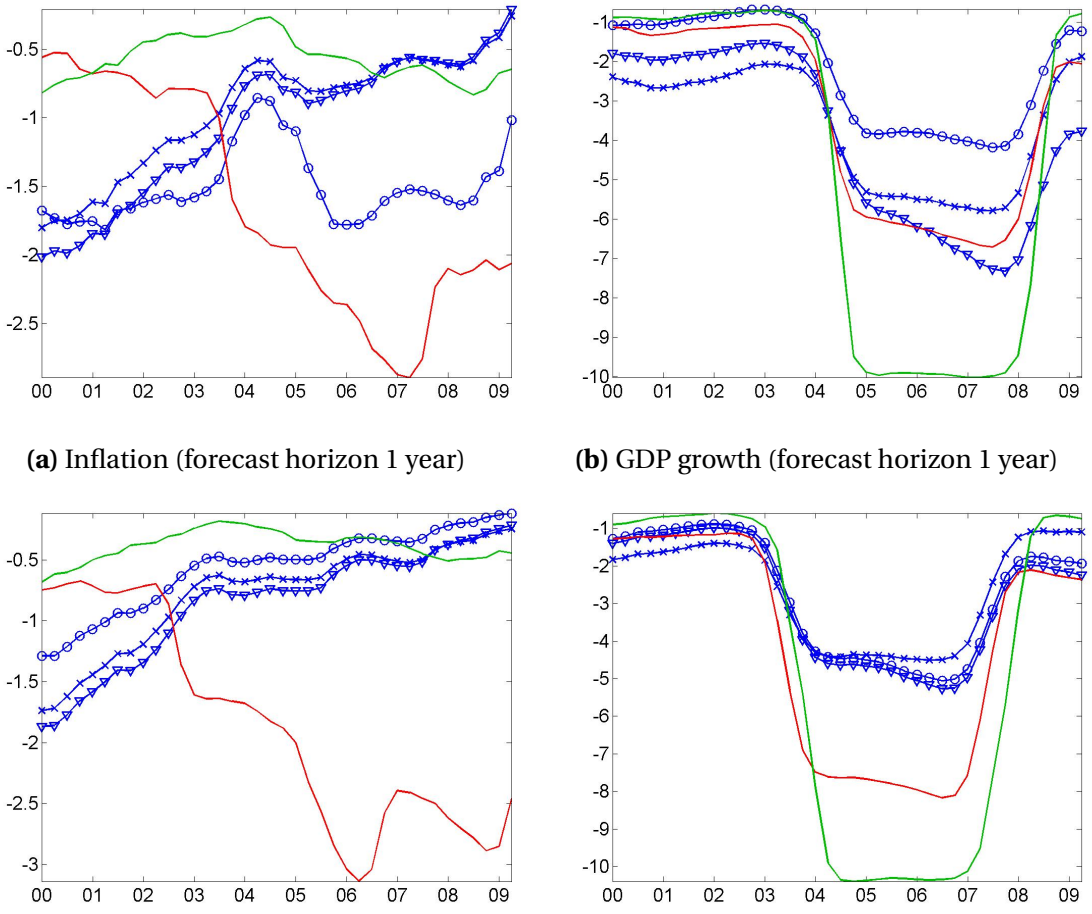
(a) Inflation in the pre-crisis period

(b) GDP growth in the pre-crisis period

(c) Inflation in the post-crisis period

(d) GDP growth in the post-crisis period

Figure 13: Average logarithmic scores for the pre- and post crisis periods at different forecast horizons.COMPASS^{PLAIN} forecasts are in blue with triangles, COMPASS^{TVT} forecasts in blue with crosses, COMPASS$^{GDP^e}$ forecasts in blue with circles, Statistical Suite forecasts in green and *Inflation Report* forecasts in red.



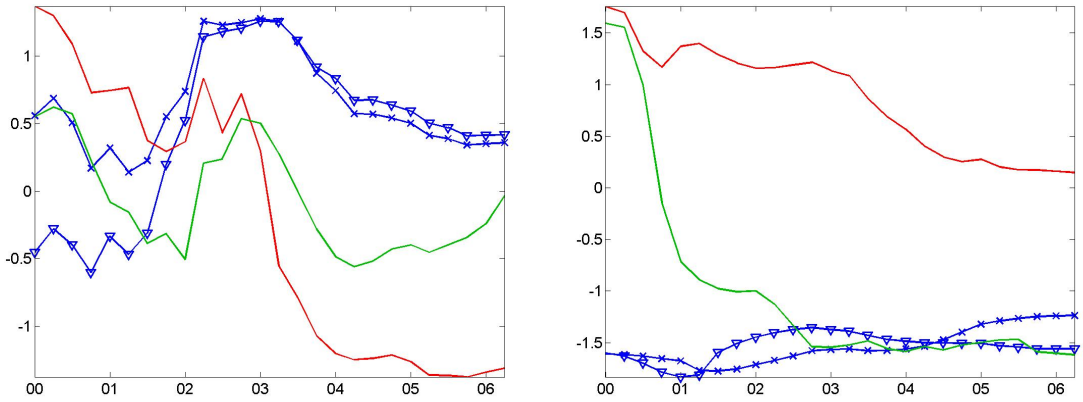
(a) Inflation (forecast horizon 1 year)

(b) GDP growth (forecast horizon 1 year)

(c) Inflation (forecast horizon 2 years)

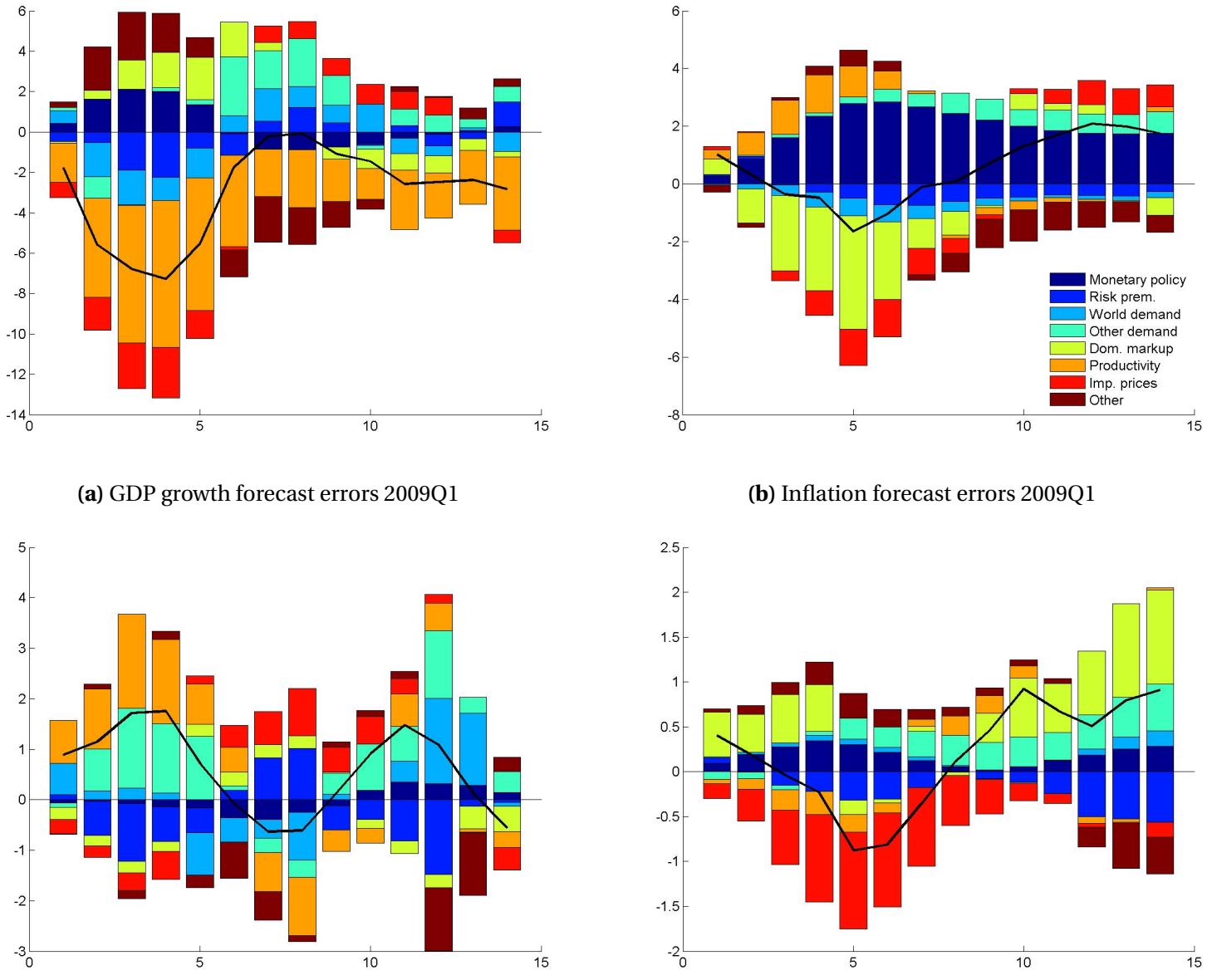
(d) GDP growth (forecast horizon 2 years)

Figure 14: Average logarithmic scores calculated over a rolling window of size 4 years at different forecast horizons. The x-axis indicates the start of the window. COMPASS^{PLAIN} forecasts are in blue with triangles, COMPASS^{TVT} forecasts in blue with crosses, COMPASS$^{GDP^e}$ forecasts in blue with circles, Statistical Suite forecasts in green and *Inflation Report* forecasts in red.



(a) Inflation **(b)** GDP growth

Figure 15: Fluctuation test statistic for a forecast horizon of 1 year based on the RMSFE. Size of the rolling window is 7 years. The x-axis indicates the start of the window. All models relative to COMPASS. A negative value of the test statistic means that COMPASS is performing better. COMPASS^{PLAIN} forecasts are in blue with triangles, COMPASS^{TVT} forecasts in blue with crosses, COMPASS$^{GDP^e}$ forecasts in blue with circles, Statistical Suite forecasts in green and *Inflation Report* forecasts in red. Critical values for a 2-sided test are 2.779 (2.5) at 5% (10%) (Table 1 in Giacomini and Rossi, 2010).



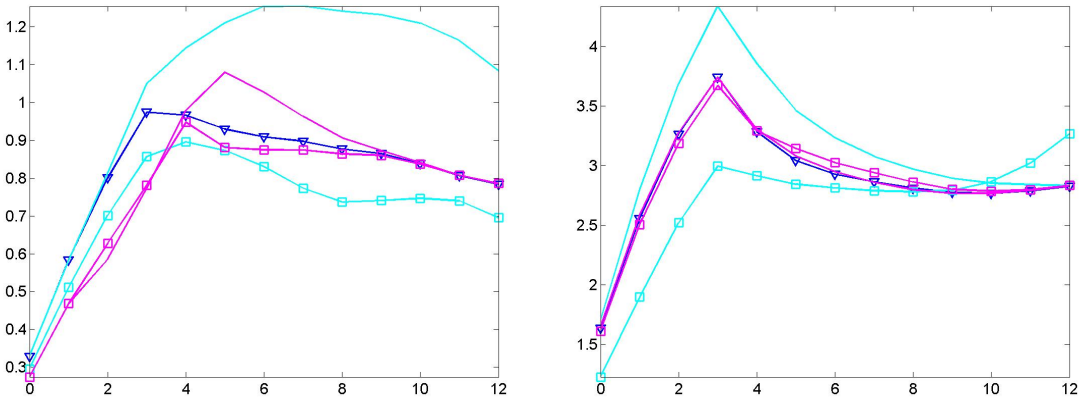
(a) GDP growth forecast errors 2009Q1

(b) Inflation forecast errors 2009Q1

(c) GDP growth forecast errors 2003Q3

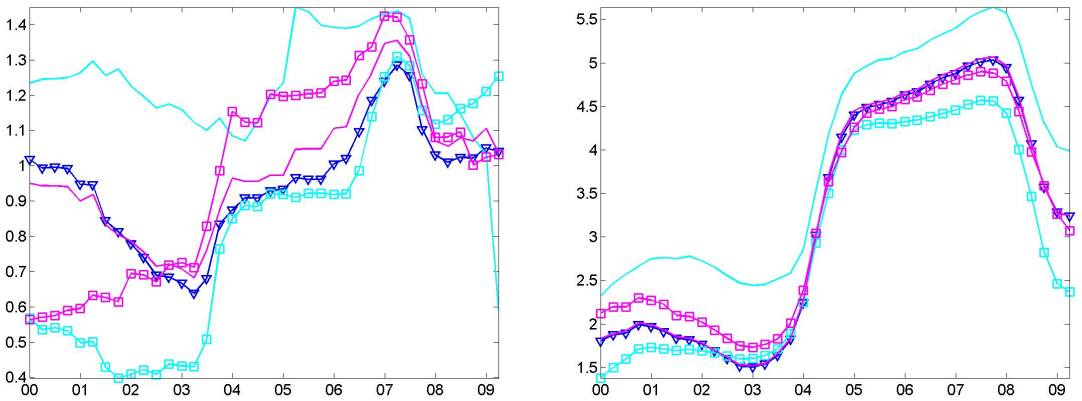
(d) Inflation forecast errors 2003Q3

Figure 16: Shock-based decompositions of the forecast errors for GDP growth and inflation based on COMPASS$^{GDP^e}$. The x axis shows the forecast horizon.



(a) Inflation **(b)** GDP growth

Figure 17: Root mean squared forecast errors for the full sample period at different forecast horizons. COMPASS^{PLAIN} forecasts without any conditioning paths imposed are in blue with triangles, COMPASS^{PLAIN} forecasts with CP^{BOE} imposed are in lightblue and COMPASS^{PLAIN} forecasts with CP^{π} imposed are in magenta. □ indicates that the conditioning paths are imposed using a full invert.



(a) Inflation **(b)** GDP growth

Figure 18: Time-varying root mean squared forecast errors for conditional forecasts at a forecast horizon of 1 year. Window length is 4 years. COMPASS^{PLAIN} forecasts without any conditioning paths imposed are in blue with triangles, COMPASS^{PLAIN} forecasts with CP^{BOE} imposed are in lightblue and COMPASS^{PLAIN} forecasts with CP^{π} imposed are in magenta. □ indicates that the conditioning paths are imposed using a full invert. The x-axis shows the start of the window.

A Grouping of shocks for shock decompositions

This appendix reports the grouping of shocks for the shock decompositions that are reported in Section 5.5:

1. Monetary policy
    - Monetary policy (14)
2. Domestic risk premium shock
    - Risk premium shock (18)
3. World demand
    - World preferences for UK exports shock (6)
    - World demand shock (16)
4. Other demand
    - Government spending shock (3)
    - Residual component of total final expenditure shock (5)
    - Investment adjustment cost shock (4)
5. Domestic price markup shocks
    - Final output price markup shock (11)
    - Value added price markup shock (12)
    - Wage markup shock (9)
6. Productivity
    - Labour augmenting productivity growth shock (2)
    - TFP shock (15)
7. Imported price shocks
    - Exchange rate risk premium shock (1)
    - World export price shock (13)
    - Import price markup shock (8)
8. Other
    - Import preference shock (7)
    - Export price markup shock (10)
    - Labour supply shock (17)

B Small-sample properties of the Diebold-Mariano test

One limitation of our study is the small sample size of 53 observations. In the case of the fluctuation test, the number of observations used to compute the test statistics is even smaller. But the small sample performance of forecast accuracy tests like the Diebold-Mariano test is poor because of the autocorrelation and heteroskedasticity consistent (HAC) covariance estimator (equation (2)) that is used to account for potential autocorrelation in the forecast errors. In this paper, we use the HAC estimator due to Newey and West (1987).[55] Because HAC estimators converge at slow, nonparametric rates, the Diebold-Mariano test can have poor properties in small samples.

To address this concern, this appendix investigates size and power of the Diebold-Mariano test in small samples by means of a Monte Carlo experiment.[56]

The data generating process is calibrated to share some properties of the forecast errors evaluated in this paper. We assume that the difference in forecast performance is generated by an AR(1) process with a persistence parameter equal to either 0.2 or 0.4, which is similar to the persistence in MSFE differences obtained from our models. To compute the empirical power, we assume that the average difference in forecast performance is either 0.5 or 1. These values are similar to what occurs in our data at forecast horizons of one or two years. The nominal size is set to 5%.

Table 10 reports the results. We find that the empirical size of the Diebold-Marino test is close to its nominal level of 5%. For sample sizes equal to 50, the empirical power under the alternative, where the differences in forecast performance is equal to 0.5, exceeds 50, even if the persistence parameter is 0.4. As expected, the empirical power is larger if the difference in forecast performance equals 1.

C 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 judgement-free, in the sense that they are unconditional forecasts, and they do not have any economic structure imposed. Rather, 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 quarter-on-quarter growth in UK GDP, and the inflation measure is quarter-on-quarter growth of the seasonally adjusted Consumer Price Index. Seasonal adjustment, where necessary, is performed using the X12 technique.

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

- 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

[55] We refer to Clark and McCracken for an exhaustive Monte Carlo study that examines the small sample behaviour of the Diebold-Mariano test with alternative HAC estimators.

[56] Because the logscore test of Amisano and Giacomini is identical to the Diebold-Mariano test except that forecast performance is measured by logscores, this appendix only reports results for the Diebold-Mariano test.

Table 10: Small sample properties of the Diebold and Mariano (1995) test

a) Empirical size

	25	50
$\rho = 0.2$	0.048	0.058
$\rho = 0.4$	0.056	0.063

b) Empirical power: Difference in forecast performance is 0.5

	25	50
$\rho = 0.2$	0.381	0.770
$\rho = 0.4$	0.179	0.509

c) Empirical power: Difference in forecast performance is 1

	25	50
$\rho = 0.2$	0.624	0.999
$\rho = 0.4$	0.287	0.942

Notes: Number of replications is 5000. The size (power) is the probability of rejecting H_0 when H_0 (H_1) is true.

- 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, whose parameters 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 3 month LIBOR interest rate, and quarterly oil price growth.
- VARM: a variant of the VAR, augmented with two variables to capture growth in money stocks
- BVAR: a variant of the VAR, estimated with Bayesian techniques
- BVARM: a variant of the VARM, estimated with 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 forecasts are combined to produce one central point forecast and probability distribution. A number of combination methods can be used, but for the exercises in this paper we have used weights given by the predictive likelihood of each model, based on a pseudo out-of-sample window of seven years.