Does Judgment Improve Macroeconomic Density Forecasts?*

Ana Beatriz Galvao
University of Warwick

Anthony Garratt
University of Warwick

James Mitchell
Federal Reserve Bank of Cleveland

This draft: January 2021

Abstract

This paper presents empirical evidence on how judgmental adjustments affect the accuracy of macroeconomic density forecasts. Judgment is defined as the difference between professional forecasters' densities and the forecast densities from statistical models. Using entropic tilting, we evaluate whether judgments about the mean, variance and skew improve the accuracy of density forecasts for UK output growth and inflation. We find that not all judgmental adjustments help. Judgments about point forecasts tend to improve density forecast accuracy at short horizons and at times of heightened macroeconomic uncertainty. Judgments about the variance hinder at short horizons, but can improve tail risk forecasts at longer horizons. Judgments about skew in general take value away, with gains seen only for longer horizon output growth forecasts when statistical models took longer to learn that downside risks had reduced with the end of the Great Recession. Overall, density forecasts from statistical models prove hard to beat.

JEL Classification: C32; C53; E37

Keywords: judgment forecasting; density forecasting; skewness; exponential tilting; forecasting uncertainty

*We thank the editor, an associate editor and two anonymous referees for helpful comments. Portions of this research were conducted under independent contract for the Federal Reserve Bank of Cleveland; the views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Cleveland or the Federal Reserve System.

1 Introduction

Professional macroeconomic forecasters and policymakers have long been understood to apply judgmental adjustments to the *point* forecasts mechanically generated by models; e.g. see McNees (1990), Fildes and Stekler (2002), Turner (1990) and Clements (1995). More recently, there has been growing interest in macroeconomics in producing density, including tail risk, forecasts; e.g. see Aastveit et al. (2019) and Adrian et al. (2019). Reflecting this, the contribution of this paper is to isolate the role of judgmental adjustments made not just about point forecasts but forecasts of the variance and skew of the distribution too. Our work complements wider inter-disciplinary evidence assessing the benefits of judgment to forecast accuracy; e.g. see Lawrence et al. (2006), Trapero et al. (2013), Davydenko and Fildes (2013) and Hyndman (2020). It also builds on work that evaluates whether the second and/or third moment forecasts from professional forecasters and central banks add value; see Knüppel and Schultefrankenfeld (2012; 2019), Kenny et al. (2015) and Clements (2018).

Judgment itself, in practice, is latent or unobserved, certainly as far as a third party observer of published macroeconomic forecasts in the UK is concerned.[1] This is because while a given professional forecaster privately knows both what if any model or models they use to guide their forecast, and the degree to which they adjust this, an outsider only observes their published forecast. At best, this observer knows something about the models or types of model which the forecaster consults. Conclusions about the role of judgment inevitably depend on relative to 'what'.

Since, in general, we neither know what model(s) is (are) used by the professional forecasters nor what judgment they apply to these, we consider density forecasts mechanically generated from a range of statistical forecasting models where no adjustments are made.[2] These models all rely on historical data and, by construction, are unable to pickup any change in the data which is expected/happening in real-time. Judgment is then taken to be the difference between the moments of the professional and statistical models' density forecast. Any observed differences

[1] In other countries and circumstances it does on occasion prove possible to identify and isolate the judgment applied to the statistical (macroeconomic) forecast. In the US, for example, the FOMC forecast (judgment-based) is published separately to the Federal Reserve Board Staff forecast (believed to be based on a statistical model). Alternatively, McNees (1990) in his assessment of the role of judgment was (privately) provided with the published (adjusted) and mechanical (unadjusted) point forecasts from four macroeconomic forecasters in the US. Similarly Clements (1995) is able to identify judgmental adjustments made to the forecasts from a leading UK forecaster given private access to their unpublished mechanical forecasts.

[2] Central bank macroeconomic forecasting systems also commonly include structural models, where judgment is incorporated via assumptions/conditional paths for exogenous variables.

we attribute to *judgment*, emphasizing that this judgment could take many forms. Judgment could involve, for example, the professional forecaster using off-model information, making assumptions about the future path of conditioning variables and/or considering alternative data and/or models to the statistical model(s).

The statistical models we consider are of differing levels of sophistication. It is well known that more complicated models do not necessarily forecast better, especially at times of structural change (Clements and Hendry, 1996). So by considering a range of models we entertain the possibility that judgment may *appear* to help when applied to a 'bad' statistical forecast but hinder relative to a 'good' model. Importantly, we consider workhorse statistical models of the type likely used and known by professional forecasters: i.e. autoregressive (AR) models, Bayesian Vector Autoregressive (BVAR) models and a density forecast combination of many statistical models. We emphasize that the density forecast combination, in particular, can generate highly asymmetric and potentially multi-modal forecast densities. As argued by McNees (1990), this evolution of the statistical model itself reflects forecaster judgment and accumulated collective learning in the forecasting profession. But mechanical use of these statistical models is as close an example to a 'pure model' forecast as one is likely to see. Their use does not require conditioning assumptions; and the forecast from the statistical model is not *post-processed* or manipulated in any way.

To help draw robust conclusions on the role of judgment relative to these statistical models, we consider the published forecasts from a range of professional macroeconomic forecasters and policymakers in the UK, all of whom are understood to rely on judgment, to some degree, to inform their forecasts. We call their forecasts "judgment-augmented". The professional forecasts considered include those from: the policymaking Monetary Policy Committee (MPC) at the Bank of England, the National Institute of Economic and Social Research (NIESR) and two surveys of professional forecasters run by Consensus Economics and the Bank of England (their Survey of External Forecasters). While an outsider can garner information about the set of structural and statistical models used by Bank of England staff to inform the MPC (see Burgess et al. (2013)), in the absence of observing both pre and post judgment forecasts, they cannot directly isolate and quantify the MPC's judgment as reflected in the fan chart forecasts that they publish each quarter in the Bank's *Inflation* (now *Monetary Policy*) *Report*. Similarly, while NIESR's structural macroeconomic forecasting model, NiGEM, is available for download to subscribers (see https://nimodel.niesr.ac.uk/), their published forecast involves unspecified

judgment and expertise being applied, as often discussed qualitatively in their forecast publication. For example, residual adjustments are made along with assumptions about the future movements of exogenous variables, such as oil prices. For the other professional forecasters, whose views are captured by the two surveys, we know even less about what, if any, models they may use to help inform their published forecasts. This explains why, to acknowledge this uncertainty, we compare their judgment-based forecasts against a range of statistical models.

Specifically, to quantify and evaluate the role of judgment, we take the density forecasts from the set of statistical models and then impose upon them specific moment(s) extracted from the density forecast of the professional forecaster, where judgment is believed to play some role. We impose this judgment systematically by exponentially tilting each of the statistical model's density forecasts to specific predictive moments obtained from the professional forecasters' density. We do this across a range of forecast horizons. Exponential tilting has also been used to add external information, including judgment-based survey forecasts, to model-based forecasts by Robertson et al. (2005), Cogley et al. (2005), Giacomini and Ragusa (2014), Altavilla et al. (2016), Krüger et al. (2017) and Tallman and Zaman (2020).

Separating the role of judgment about the mean from higher moments is important and central to our contribution. Professional forecasters increasingly convey their assessments about the future values of macroeconomic variables not only by expressing the most likely values of these variables (the point forecasts) but by also communicating the uncertainty around these central predictions. Indeed, to reflect differing balances of positive and negative risks, they publish asymmetric density forecasts with skew (Britton et al., 1998). Accordingly, in this paper, we evaluate judgments made not only about the first moment of the predictive density (as in Altavilla et al. (2016)) but also the second (as in Cogley et al. (2005), Krüger et al. (2017), Clements (2018) and Tallman and Zaman (2020)) and the third moment (forecast skewness) arising from the perception of asymmetric risks.

We consider forecasts for year-on-year quarterly GDP growth and inflation in the UK for quarterly forecasting horizons up to two years ahead. To ensure no look ahead bias, we use real-time data vintages on GDP growth. We consider different loss functions to measure the impact of judgment on different aspects of density forecast accuracy. These include the threshold-weighted CRPS (Gneiting and Ranjan, 2011) that allows us to concentrate evaluation on tail events.

We conclude that judgmental adjustments do not, in general, tend to improve the accuracy

of density forecasts from statistical models. But there can be gains in specific instances. Judgment, about the mean, improves short-horizon density forecasts especially for output growth during turbulent periods, such as the financial crisis (2008/2009). In contrast, judgmental assessments about forecast uncertainty significantly diminish density forecast accuracy at short horizons, in particular during the relatively stable 2013-2016 period. At long horizons, these judgmental adjustments can improve the accuracy of tail event forecasts, in particular when applied to statistical models that do not accommodate time-variation in the conditional variance. We also find that judgments about skew - about the balance of risks – typically lower the forecasting accuracy of statistical models. But skew judgments can improve tail risk forecasts for output growth at times of macroeconomic change, such as around business cycle turning points. Statistical models took longer to learn that downside risks had reduced with the end of the Great Recession.

The plan of the remainder of this paper is as follows. Section 2 details the four professional forecasters and the three statistical forecasting models. Section 3 presents the moments extracted from the judgmental-augmented forecasts and the statistical models. Section 4 explains how the density forecasts from the statistical models are tilted to satisfy moment conditions from the judgment-based density forecasts of the professional forecasters. The loss functions used to measure forecast performance are described. Section 5 presents the empirical results. It considers if, how and when time-varying judgments about the mean, variance and skew of output growth and inflation can inform the density forecasts produced by the statistical models. Section 6 concludes. Online appendices contain additional details and supplementary empirical results.

2 Professional Forecasters and Statistical Forecasting Models

In this section, we summarize the judgment-augmented and statistical forecasting models. For details, see online Appendices A and B, respectively.

We consider four sets of professional forecasts. These are forecasts from: (i) NIESR; (ii) the MPC at the Bank of England; (iii) the Survey of External Forecasters (SEF), run by the Bank of England; and (iv) Consensus Economics (CE). The forecasts that we analyze from both SEF and CE in fact involve aggregating the forecasts from a far larger set of professional forecasters whose opinions are sought by the respective survey.

We consider three statistical forecasting models, including a combination of statistical models as defined in the Warwick Business School Forecasting System (WBSFS).

2.1 Professional Forecasters

For NIESR, we focus on forecasts published in their quarterly publication - the *National Institute Economic Review*. NIESR started producing density forecasts for GDP and inflation in 1996, quantifying forecast uncertainty based on historical point forecasting errors. In 2008 they switched to producing their densities via stochastic simulation. But, unfortunately, the parameters of these forecasts have not always been tabulated in the *Review*. As a consequence, it is not possible to construct a consistent historical time-series of their density forecasts as computed each quarter.[3] Therefore, in this paper, we use their point forecasts only. Specifically, we consider historical data on NIESR's point forecasts from 1992Q1 for GDP growth and from 2002Q1 for CPI inflation.

The MPC density forecasts for GDP growth and inflation are two-piece normal densities. The two-piece normal creates a potentially skewed density by combining the two halves of two normal densities with a common mode of μ and standard deviations of σ_1 and σ_2. In our application, we use the first three moments of the two-piece normal, as published in real-time by the Bank of England.[4] These moments, for the random variable Y, are given as:

$$E(Y) = \mu + \sqrt{\frac{2}{\pi}}(\sigma_2 - \sigma_1) \quad (1)$$

$$var(Y) = \left(1 - \frac{2}{\pi}\right)(\sigma_2 - \sigma_1)^2 + \sigma_1\sigma_2 \quad (2)$$

$$skew(Y) = \frac{\sqrt{\frac{2}{\pi}}(\sigma_2 - \sigma_1)\left(\frac{4}{\pi} - 1\right)(\sigma_2 - \sigma_1)^2 + \sigma_1\sigma_2}{var(Y)^{3/2}} \quad (3)$$

where we use the Pearson moment coefficient of skewness, (3). We consider CPI inflation forecasts from 2004Q1, when the targeted measure of inflation switched from RPIX to CPI inflation.

For the SEF, we focus on the aggregated histograms. These are the average probabilities, across forecasters, given to realized output growth or inflation lying within a set of pre-assigned

[3] Prior to 2008 NIESR's density forecasts were assumed Gaussian. They were centered on the point forecast published in the *Review*; and the predictive variance can be inferred from the standard deviations or root mean square forecast error estimates that used to be published; see Mitchell (2005) for details.

[4] Spreadsheets containing the parameters for all the published fan charts are publicly available on the BoE's web site (http://www.bankofengland.co.uk/publications/inflation nreport.irprobab.htm).

intervals or bins. Average or combined density forecasts are commonly used to represent the consensus opinion of forecasters. The average forecast produced by the SPF in the US is a notable example.[5] As reviewed by Aastveit et al. (2019), combined densities are also often found to work well empirically. Following Engelberg et al. (2008) and Clements (2014), we fit a generalized-beta distribution, which allows for asymmetry, to these aggregated histograms. From these fitted densities we then extract mean, variance and skewness forecasts.[6] We focus on SEF forecasts from 2006Q2, as this is when one and two year ($h = 4$ and $h = 8$ quarter) ahead "fixed-horizon" questions were introduced for both output growth and CPI inflation.[7]

For CE, we consider the average of their 30-40 survey participants' point forecasts. Timmermann (2006) and others have shown the effectiveness of combined point forecasts. We consider quarterly CE point forecasts for horizons from one through seven quarters ahead for output growth and CPI inflation from 1999Q1.

2.2 Statistical Forecasting Models

For our main benchmark statistical model, we use a combination of statistically-motivated econometric models common in the literature, as implemented in the Warwick Business School Forecasting System (WBSFS).[8] The emphasis is on a combined model – as opposed to a single model - which produces judgment-free point and density forecasts. As reviewed in Aastveit et al. (2019), a combination of forecast densities has been found to be an effective means of accommodating "model uncertainty" when forecasting; moreover, professional forecasters most likely consult a wide variety of models. The use of a combination can reflect the fact that all models are likely misspecified, such that relative (across model) forecast performance changes over time. Structural breaks, in particular, are well known to contribute to the unreliability of point and density forecasts (e.g., see Clements and Hendry (1996)) and lead to instabilities in the performance of specific forecasting models. As, e.g., Hendry and Clements (2004) and Pesaran and Timmermann (2007) explain, models differ in their sensitivity to structural breaks.

[5] The SPF in the US takes a linear combination of the individual density forecasts and publishes mean and median point forecasts.

[6] As a robustness check, we also calculated the moments directly (non-parametrically) from the histograms. The results are discussed in Appendix C.1.

[7] Fixed-horizon event questions for two-year ahead output growth were introduced in 1998; and in May 2006 for both CPI inflation and output growth for 1, 2 and 3 year ahead horizons. Note also that the original inflation based questions were about RPIX inflation.

[8] The WBSFS forecasts were published quarterly between November 2014 and October 2019 on the WBS website (https://warwick.ac.uk/fac/soc /wbs/subjects/finance/mpf/forecasting) and in the *National Institute Economic Review* between October 2017 and October 2020.

As a result, combining density forecasts across different types of model offers the promise of more robust forecasts. Importantly, density forecast combinations can generate densities with different characteristics to the component densities. By differentially weighting the component models, density combinations can also accommodate time-varying volatility, even when models directly modeling volatility are not being combined. Jore et al. (2010), e.g., show how combinations of densities from a range of AR and VAR models pickup the decline in volatility associated with the Great Moderation. In our case, given heightened awareness of the need to model volatility, we also consider a BVAR with stochastic volatility as a component model.

The WBSFS combines the densities of 24 forecasting models from the following three *classes* of statistical model: (i) quarterly BVARs, (ii) mixed-frequency models that also exploit monthly data and (iii) quarterly autoregressive models. Here we describe the model classes in general terms only. Details are provided in Appendix B.1.

Within the first class of model, three types of BVAR are considered. These BVARs differ according to the (quarterly) variables considered and whether stochastic volatility is modeled. First, a single BVAR models the log-levels of the seven variables used in Smets and Wouters (2007) DSGE model, described in Appendix B.1. Second, is a set of medium-sized BVARs which, in addition to GDP and CPI inflation, include thirteen indicators, as described in Appendix B.1 and Table A6. We estimate eight variants of this medium-sized BVAR, for lag lengths $p = 1, \ldots, 4$, in both log-levels and first-differences of the data. Finally, we consider a single medium-sized BVAR with stochastic volatility in these 15 variables. There is evidence that accommodating stochastic volatility is especially helpful when density forecasting; see Clark (2011). In total, the BVAR model class comprises 10 models.

The second class of model comprises mixed-frequency models. These allow for consideration of known (within-quarter) monthly information, reflecting their release calendars. Specifically, Autoregressive Distributed Lag Mixed Data Sampling (ADL-MIDAS) models are used to relate the quarterly target variable (GDP growth or CPI inflation) to each of the aforementioned thirteen monthly indicators, described in Table A6. Finally, for the third class of model, we consider an AR(2) model. In total, the WBSFS therefore combines 24 individual models.

For robustness, and to acknowledge that judgment may *appear* to help when applied to a 'bad' statistical forecast but hinder relative to a 'good' model, we also compare the professional forecasts against those from both a small BVAR model (in the first differences of GDP, CPI, the unemployment rate, the three-month interest rate and the real effective exchange rate index),

as is standard in the tilting literature (Robertson et al., 2005; Krüger et al., 2017), and an AR model. They are described in Appendix B.2 and B.3. For space reasons, we focus in the main paper on reporting results for the WBSFS and discuss the AR and BVAR results only when this affects inference.

3 Characteristics of Judgment-Augmented Forecasts

While the characteristics of judgment-augmented point forecasts for UK output growth and inflation have been discussed previously (e.g. see Turner (1990) and Clements (1995)), less is known about how judgment impacts forecast variances and skewness. So, in this section, we characterize the temporal variation in the judgment-augmented variance and skew forecasts. We compare them against the WBS combination density forecasts.

3.1 Types of Judgment

For each of the professional forecasters described in section 2.1 we extract, at each forecast origin, τ, and for each forecast horizon, h, information on the first three moments of their density forecasts: the mean, standard deviation and (standardized) skewness. Each of our four professional forecasters (NIESR, MPC, SEF and CE) provides point forecasts for output growth and inflation. But, as summarized above, availability of higher moment information is more limited. Only the MPC and the SEF have consistently communicated the uncertainty (defined here as the standard deviation) around their forecasts. Given that their density forecasts need not be symmetric, for the MPC and the SEF we also extract their estimates of predictive skew.

3.2 Time-variation in the forecasts for uncertainty and skew

Figure 1 plots the evolution over time of the judgment-adjusted forecasts of uncertainty (standard deviation) for output growth and inflation at three forecast horizons. For comparison, the uncertainty forecasts from the WBS combination are also shown.[9] The figures are drawn with the x-axis indicating the forecast origin.

Figure 1 reveals sharp and pronounced movements in the MPC's uncertainty forecasts, especially at shorter horizons. Their one-quarter-ahead uncertainty forecasts, for both output growth and inflation, have more than tripled in size. For output growth, these increases in

[9] We do not plot the mean forecasts because all four professionals produce quite similar output growth point forecasts.

predicted uncertainty are sharp and large, particularly in 2008-9 at the time of the global financial crisis. For inflation, the MPC's uncertainty forecasts increase more smoothly between 2006 and 2011, but since then have remained fairly constant. By contrast, for both one-year-ahead and two-year-ahead uncertainty (middle and lower panels of Figure 1, respectively), the uncertainty estimates from the SEF are much lower than those reported by the MPC.

It is also of interest that the MPC's uncertainty forecasts do not reflect historical point forecasting performance. As discussed in Appendix C.2, the MPC's variance forecasts exceed the variance of their past forecast errors. This suggests that the MPC do indeed apply judgment when setting the variance of their fan chart.

Comparison against the uncertainty forecasts from the WBS combination reveals sensitivity to the forecast horizon. One-quarter-ahead, the MPC predicted two-to-three times more uncertainty than the WBS combination. But their assessments of uncertainty are more similar at longer horizons. The judgment-augmented uncertainty forecasts made by the MPC and the SEF increase for both output growth and inflation during 2008-2009; whereas the WBS combination uncertainty forecasts are more stable, especially for inflation.

Figure 2 presents the predicted skew forecasts by forecast origin and forecaster. Figure 2A shows the MPC's skew forecasts. The MPC has always viewed risks to output growth to be on the downside. For inflation, predicted skew fluctuates between the dominant risk being on the upside and the downside. Positive shocks were seen as more likely in some periods, such as 2010-2011, and downside risks in others, such as 2015-2016.

Figure 2B compares the two-years-ahead MPC and SEF assessments of predicted skew to the equivalent forecasts from the WBS combination.[10] Figure 2B reveals that MPC assessments of the balance of risks to inflation differ to those from the statistical model. For output growth, there is more agreement - both the MPC and the WBS combination predict negative skew in 2008-2009. The judgment-augmented skew forecasts from the MPC react especially sharply from late 2008 to the emerging downside risks; they also return to a more balanced assessment quicker than the statistical model. There is more persistence to the statistical model's skew forecasts. The SEF skew forecasts for output growth are always negative, while their skew forecasts for inflation fluctuate in sign. Again, we see the SEF skew forecasts differ from those from the statistical model. This evidence of disagreement between the skew forecasts from

[10] As expected, given underlying symmetry assumptions, the skew forecasts from the AR and the BVAR models are both smaller and less variable, over time, than those from the WBS combination; see Figure A6.

professional forecasters and statistical models suggests that MPC and SEF forecasts of skew are, at least in part, judgment-based - with experts making different judgments.

4 Empirical Methodology

To evaluate the impact of judgment when forecasting, we exponentially tilt, using the methodology explained in section 4.1 below, the statistical models' density forecasts so that they satisfy the mean, standard deviation and/or skew judgment-augmented forecast. We then measure and evaluate the effects of the judgment-based tilted density forecasts using the loss functions and statistical tests described in section 4.2.

4.1 Tilting forecasts using judgment-based moment restrictions

Exponential tilting, as introduced into macroeconomic forecasting by Robertson et al. (2005), involves modifying a given predictive distribution into a new predictive distribution to satisfy a set of moment conditions or restrictions, but minimizing the relative entropy or distance between the two distributions.

Our objective, for each forecasting origin in the out-of-sample period $\tau = T+1, ..., T+P$, is to construct from the statistical model's density forecast, $f_{\tau,h}(Y)$, a new density forecast, $\widetilde{f}_{\tau,h}(Y)$, using data up to and including period τ, that satisfies judgment-based moment conditions.

Specifically, draws are first taken from the predictive density, $f_{\tau,h}(Y)$, produced from one of the statistical models described in section 2.2. Then, recursively for each τ, we tilt $f_{\tau,h}(Y)$ to satisfy a set of k moment conditions:

$$E_\tau[g_{\tau,h}(Y) - \bar{g}] = 0 \tag{4}$$

where $[g_{\tau,h}(Y) - \bar{g}]$ is a $k \times 1$ vector of moment conditions, which we will define for the mean, variance, skew and/or kurtosis. Robertson et al. (2005) and Giacomini and Ragusa (2014) show that tilting delivers a new density forecast, $\widetilde{f}_{\tau,h}(Y)$, that satisfies (4) but in a manner that keeps it as *close* as possible to the original density, $f_{\tau,h}(Y)$, according to the Kullback-Leibler measure of distance.

The tilted density $\widetilde{f}_{\tau,h}(Y)$ is computed as:

$$\widetilde{f}_{\tau,h}(Y) = f_{\tau,h}(Y) \exp\left\{\eta_{\tau,h} + \tau'_{\tau,h}(g_{\tau,h}(Y) - \bar{g})\right\}, \tag{5}$$

where $\eta_{\tau,h}$ and $\tau_{\tau,h}$:

$$\tau_{\tau,h} = \arg\min_{\tau} \frac{1}{D} \sum_{d=1}^{D} f_{\tau,h}(y^d) \exp\left\{\tau'(g_{\tau,h}(y^d) - \bar{g})\right\} \quad (6)$$

$$\eta_{\tau,h} = \log\left\{\frac{1}{D} \sum_{d=1}^{D} f_{\tau,h}(y^d) \exp[\tau'_{\tau,h}(g_{\tau,h}(y^d) - \bar{g})]\right\}^{-1}, \quad (7)$$

and $\left\{y^d\right\}_{d=1}^{D}$ are $D = 10,000$ draws from $f_{\tau,h}(Y)$.[11]

To help understand the role of judgment about specific moments of the forecast distribution, we tilt in two ways. The first approach ("Tilting Approach 1") involves imposing, via (4), one judgment-augmented moment restriction at a time while keeping the other moments fixed at their values from the statistical model. The second approach ("Tilting Approach 2"), as more conventionally used in the tilting literature, imposes moment restrictions jointly and does not force other moments to remain at their values from the statistical model.

In more detail, Tilting Approach 1 tilts separately towards either the mean, the variance or the (Pearson moment) skewness of the judgment-based forecast.[12] The k vector of moment conditions, $\bar{g}$, in (4), is set equal to the first four moments of $f_{\tau,h}(Y)$ except for the specific moment that is tilted towards the judgment-based forecast. This form of tilting is, in effect, a constrained minimization, with only higher-order moments (beyond the k-th) free to adjust. We note (e.g. see Giacomini and Ragusa (2014)) that if $f_{\tau,h}(Y)$ is Gaussian, tilting towards a different mean forecast also involves keeping other moments unchanged. As proven in Giacomini and Ragusa (2014), if the judgment about the specific moment that we tilt towards is true in population, we should expect the tilted density forecast $\widetilde{f}_{\tau,h}(Y)$ to outperform (according to the logarithmic scoring rule evaluated at the subsequent realization, y_{t+h}) the original density forecast $f_{\tau,h}(Y)$. For robustness, and aware that kurtosis estimates can be imprecisely defined, we also consider constraining only the first $k = 3$ moments.

While Tilting Approach 1 isolates the effects of judgment on specific moments, it does so assuming that forecasters' judgments about one moment are independent of the others. This may or may not be a reasonable assumption for what are in large-part subjectively formed densities. That is, while statistically the (conditional) variance forecast, for example formed via OLS estimation of an AR model, depends on the (conditional) mean forecast (under mean

[11] See Giacomini and Ragusa (2014) for details. Further computational details are in Appendix C.3.

[12] Since Tilting Approach 1 is unfamiliar relative to Tilting Approach 2, we illustrate how it works in practice in Appendix D. The illustration shows how Tilting Approach 1 affects the shape of the forecast densities made during the Great Recession.

squared error loss), professional forecasters may or may not form their density forecasts in this *holistic* manner. To acknowledge that they might, in Tilting Approach 2 (as in Krüger et al. (2017)) we tilt towards: i) the judgment-based mean forecast ($k = 1$); ii) the mean and variance of the judgment-based forecast ($k = 2$); and iii) the mean, variance and skewness of judgment-based forecast ($k = 3$).[13] Note how Tilting Approach 2 no longer forces the other moments to remain at their values from the statistical model. Incremental comparisons of forecast accuracy between i), ii) and iii) let us evaluate if additional judgment-based moments improve forecast accuracy. For robustness, we also consider imposing just one moment (whether it be the mean, variance or skewness) at a time, i.e. we set $k = 1$; but again, in contrast to Tilting Approach 1, the other moments are left unconstrained.

4.2 Measuring Forecasting Performance

4.2.1 Loss Functions

To evaluate the accuracy of the density forecasts we use the continuous ranked probability score (CRPS) and the threshold-weighted CRPS. Both are loss functions, $L(f_{t,h}(Y), y_{t+h})$, that score the density forecast, $f_{t,h}(Y)$, according to the realization, y_{t+h}, that subsequently materializes (Gneiting and Ranjan, 2011). $L(.)$ is defined so that smaller values indicate greater accuracy. The CRPS evaluates the 'whole' density, while the threshold-weighted CRPS focuses on accuracy in the tails.[14]

Specifically, the CRPS is given as:

$$CRPS_{t,h} = \int_{-\infty}^{+\infty} \left[F_{t,h}(y) - I(y_{t+h} \leq y)\right]^2 dy \tag{8}$$

where $F_{t,h}(.)$ is the CDF associated with the density forecast $f_{t,h}(.)$ and $I(y_{t+h} \leq y)$ denotes an indicator function equal to unity if $y_{t+h} \leq y$, 0 otherwise.

The threshold-weighted CRPS is:

$$twCRPS_{t,h} = \int_{-\infty}^{+\infty} w(y) \left[F_{t,h}(y) - I(y_{t+h} \leq y)\right]^2 dy \tag{9}$$

[13] We refer to Krüger et al. (2017) for some illustrative examples of how this second form of tilting affects the shape of the forecast distributions.

[14] Appendix C.4 provides computational details for the CRPS. As robustness check, we also consider the logarithmic score.The logarithmic score is less robust to outliers than the CRPS. But results are qualitatively similar to those using the CRPS.

where $w(y)$ are positive weights. $twCRPS_{t,h}$ is a *proper* scoring function; cf. Lerch et al. (2017). When $w(y) \equiv 1$ for all y, $twCRPS_{t,h}$ reduces to the unweighted CRPS, (8). Otherwise, $w(y)$ can be tailored to focus on specific regions of the density. We specify $w(y)$ to focus on the tails. If we define a Gaussian CDF for y as $\Phi(y, mean, var)$ and let the two thresholds, r_1 and r_2, define regions in the left and right tails, then $w(y)$ is set as:

$$w(y) = (1 - \Phi(y, r_1, var)) + \Phi(y, r_2, var). \tag{10}$$

As a consequence, $twCRPS_{t,h}$ evaluates differences between the predicted and true CDF only for regions of the density below r_1 and above r_2. How clearly these regions are defined depends on var; we set $var = 0.2$, so that the weights move sharply to 1 as y crosses either threshold. Consulting historical data for output growth and inflation from 1980 to 2016, we set the thresholds so that approximately 10% of realizations fall in each tail. For GDP growth, this involves setting $r_1 = 0\%$ and $r_2 = 4\%$; for inflation, $r_1 = 1\%$ and $r_2 = 4\%$. As a consequence, $twCRPS_{t,h}$ favors density forecasts best able to characterize tail events.

4.2.2 Statistical Tests

To test for statistically significant differences in forecast accuracy between the original density, as evaluated by the chosen loss function, $L(f_{t,h}(Y), y_{t+h})$, and the tilted density, $L(\tilde{f}_{t,h}(Y), y_{t+h})$, we test the null of equal forecast accuracy:

$$H_0 : E[L(\tilde{f}_{t,h}(Y), y_{t+h}) - L(f_{t,h}(Y), y_{t+h})] = 0. \tag{11}$$

Following Diebold and Mariano (1995) and Giacomini and White (2006), we use a t-statistic assuming asymptotic normality and implement a two-sided test.[15] We reject the null in favor of the tilted distribution when the t-statistic is negative and smaller than the critical value. Following Harvey et al. (2017), we use a rectangular kernel with the lag truncation parameter set to $h - 1$ to obtain HAC standard errors, and apply a small sample correction.[16]

This test is computed over P observations in the out-of-sample period. However, we are also interested in whether there are changes in forecast accuracy over these P observations; and if there are significant differences in accuracy between the original and the tilted densities when

[15] The impact of decreasing parameter uncertainty due to recursive estimation is assumed to be negligible, motivated by Diebold and Mariano (1995) and Diebold (2015).

[16] We thank a referee of this journal for suggesting this approach.

we allow for changes in relative accuracy. Accordingly, we make use of the fluctuation test of Giacomini and Rossi (2010). This involves computing the t-statistic over rolling windows within the out-of-sample period, but with the long-run variance computed over all P observations. We set the window size to 20, which is about $1/3P$ for output growth, and $1/2P$ for inflation.

5 Assessing the Impact of Forecasters' Judgment

5.1 Design of the Empirical Exercise

We use real-time data for real GDP, consumption and investment from the ONS. UK CPI inflation is not subject to data revisions. Due to a lack of availability of vintage data, we ignore revisions for some of the additional indicator variables in the WBS combination, such as industrial production and real wages; but we do take into account their publication delays, i.e. we use values as available in the middle month of each quarterly forecast origin.

All models are re-estimated at each forecast origin τ, using an expanding window of data from 1980Q1 (1980M1 for models that use monthly data) through $\tau - 1$, where $\tau = T + 1, ..., T + P$.[17] The forecast origins date from 2001Q1 through 2016Q1 for output growth; and from 2004Q1 through 2016Q1 for inflation. The latest data vintage that we consider is for 2018Q2, so that the same number of observations is used to evaluate the forecasts across the $h = 1, ..., 8$ quarter-ahead forecast horizons. This results in $P = 61$ for output growth and $P = 49$ for inflation.

Given GDP data revisions, we have to decide which vintage of data to use to define the *realization,* y_{t+h}. We use the realization for year-on-year GDP growth published by ONS two months after the end of the reference quarter. Over our sample period, this is the second GDP release.

5.2 The effect of judgment on forecast performance

5.2.1 Tilting Approach 1: imposing one moment at a time

Table 1 evaluates the role of judgment about the mean, variance and skew forecast. Recall, Tilting Approach 1 involves tilting the statistical model's density forecast, at a given horizon, to a specific judgment-based moment forecast. But the remaining $k - 1$ moments (up to $k = 4$

[17] Given sample sizes are fairly small, we use expanding rather than rolling windows. Preliminary results suggested that this improved forecast accuracy.

in (4)) are constrained to their original values. We focus here on use of the WBS combination density as the statistical density forecast. Appendix Tables A1 and A2 present results for the AR and BVAR statistical model, which we summarize below. We emphasize that no single statistical model consistently delivers the lowest CRPS across variables and horizons.

The first column of Table 1 reports the average CRPS and tw-CRPS statistics from applying Tilting Approach 1. The entries in the remaining columns of Table 1 report the ratio of the score statistic of the tilted density to the statistical density (WBS combination). Ratios less than one indicate improvements in forecasting accuracy due to the imposition of judgment. Entries in bold indicate rejection of the null hypothesis of equal forecast accuracy in favor of the tilted density that incorporates judgment. In contrast, underlined entries indicate rejection in favor of the statistical density, implying that judgment worsens forecast accuracy.

Panels A (output growth) and B (inflation) show results for horizons $h = 1, 4$ and 8. Panel C reports proportions over all 8 forecast horizons, variables (output growth and inflation) and score functions (CRPS and tw-CRPS). Specifically, the rows report the proportion of times that judgment improves density forecasting, or has either significantly improved or worsened performance at the 90% significance level (according to individual t-tests).

Our overriding interest is drawing out whether and how judgments about the mean, variance and skew improve forecast accuracy. We are not interested *per se* in establishing whether, *ex post*, one professional forecaster (or group of) is better than another. We take the view that *ex ante* it is hard for an independent observer to know which professional forecast is best. What matters in our exercise is whether *collectively* judgment adds value to the density forecasts produced mechanically by models. Hence we do not emphasize individual professional forecaster performance.

Accordingly, looking at Panel C, the main takeaway from Table 1 is that while judgments about the mean improve forecasting performance on 69% to 88% of occasions (these improvements are statistically significant on 3% to 25% of occasions), judgments about the variance hinder. Forecast performance improves after imposing judgment about the variance only on 38% to 50% of occasions. There is even less value-added to the judgment-based skew forecasts.

These results tend to be robust to consideration of the logarithmic scoring rule instead of the CRPS; robust to considering the AR and BVAR densities as the statistical density forecast; and robust to tilting up to $k = 3$, so that only the first 3 moments are constrained in Tilting

Approach 1[18]; see Appendix Tables A1 through A4. The only qualification to this is that the judgment-based variance forecasts do improve accuracy, especially at longer horizons, relative to the AR and BVAR statistical models on up to 70% of occasions. This is consistent with the AR and BVAR forecasts tending to forecast the tails, especially for output growth, less accurately than the WBS combination (compare the first column of Table 1 with Tables A1 and A2), perhaps as they do not accommodate time-variation in the conditional variance like the WBS combination.

5.2.2 Tilting Approach 2: imposing multiple moments

Table 2 follows Table 1, but evaluates judgments about the mean, variance and skew forecast using Tilting Approach 2. Consistent with Table 1, we find that tilting towards the mean and then separately towards the variance improves and worsens, respectively, forecast performance. Mean judgments deliver improvements on 75% to 88% of occasions (and are statistically significant 13% to 63% of the time). But variance judgments only deliver gains on 16% to 38% of occasions (with 0% to 38% being statistically significant). In contrast, tilting towards the judgment-enhanced predicted skew alone now has a more beneficial effect with improvements in 38% and 63% of the cases, relative to 13% to 34% under Tilting Approach 1. Interestingly, skew judgments improve the forecasting performance more frequently when tail performance is emphasized via the tw-CRPS. Similar results are found when the BVAR is considered as the statistical model (cf. Table A2 and Table A5).

Looking next at tilting towards the mean, the mean and variance, and then jointly towards all three judgment-based forecasts, Table 2 reveals no incremental gains relative to tilting towards the mean only. That is, while tilting towards multiple judgment-based moments delivers gains on up to 88% of occasions, these gains materialize even when tilting just towards the mean. We therefore conclude that results are robust across Tables 1 and 2: mean judgments help most.

5.3 The Time-Varying Effects of Judgment Adjustments

We use the fluctuation test described in section 4.2.2 to assess whether our general finding that, on average over the evaluation period, judgment about the central forecast (in particular for output growth) helps and that judgment about higher moments hinders, in fact masks temporal

[18] As emphasized by a referee, kurtosis estimates can be imprecise. Hence, for robustness, we considered Tilting Approach 1 but not constraining the kurtosis of the tilted density to equal the value from the statistical model.

variation. We continue to focus on the WBS combination as the statistical benchmark.

In Figure 3 we present t-statistics from the fluctuation test computed over rolling windows of 20 quarters. Negative values indicate that judgment improves the accuracy of the WBS combination: values less than -2.8 indicate statistically significant improvements at the 10% level (greater than 2.8 indicates statistically significant losses).[19]

Figure 3.1 shows that the density forecasting gains seen in Tables 1 and 2 to tilting the one-year-ahead output growth density forecasts to the mean forecasts from the professional forecasters (for space, we focus on the judgment of the MPC and CE) largely arise during the turbulent 2009 to 2012 Great Recession period. Judgment about the central path appears to add value at times of change.

In contrast, Figure 3.2 reveals that the losses in density forecasting accuracy seen one-quarter-ahead when imposing judgment on uncertainty arise primarily in the period since the Great Recession. Consistent with Figure 1, this appears to reflect the MPC not lowering the variance of their short horizon inflation and output growth density forecasts, after raising their variance forecasts so sharply during the Great Recession.

While we have found judgments about the variance to hinder at short horizons, Table 1 did indicate that these judgments can help deliver better tail risk forecasts, especially for output growth, at longer horizons. To investigate further, Figure 3.3 evaluates the time-varying contribution of MPC forecasts of two-year-ahead uncertainty. Figure 3.3 indicates that these judgments improve tail and density forecast accuracy for output growth from 2014 onwards. During this period, as seen in Figure 1, the uncertainty predicted by the WBS combination was slowly declining. But the width of the MPC density forecast for output growth contracted quite sharply in 2013: and Figure 3.3 suggests that this judgment helped, even if, on average over the evaluation period, the effects of tilting the WBS combination towards the MPC variance forecast are more modest (cf. Tables 1 and 2). There is no evidence from Figure 3.3 that judgments about inflation uncertainty really helped, even at specific points in time.

Figure 3.4 investigates if there were periods in time when skew judgments really helped. We focus on the one instance detected in Tables 1 and 2 when skew judgments did, on average over time, help - when forecasting two-years-ahead. Figure 3.4 confirms that the gains seen tilting the WBS combination density forecast for output growth to the MPC skew forecast are largely

[19] The critical values are lower for inflation as the windows of 20 observations are a larger proportion of the sample size. The relevant critical value is then 2.5. Critical values are obtained from Giacomini and Rossi (2010).

confined to 2010-2011, when using the tw-CRPS to isolate tail forecast accuracy. This period is when (see Figure 2B) the WBS combination forecast more skew than the MPC, given that the MPC sharply reduced its assessment of downside risks knowing, unlike the statistical model, that the recession had ended. We again interpret this as evidence that judgments about skew can improve probabilistic forecasts, especially at times of macroeconomic change. However, we emphasize that overall skew judgments rarely improve the density and tail risk forecasts from statistical models. Indeed, Figure 3.4 shows, for inflation, that since 2014 the MPC's skew judgment has led to worse density forecasts.

6 Conclusions

This paper presents empirical evidence on the role that judgment plays in macroeconomic density forecast accuracy. In an application to UK output growth and inflation, we find that density forecasts from statistical models prove hard to beat. Only selected judgmental adjustments improve statistical models' density forecasts.

Judgments about the mean improve density forecast accuracy at short horizons, especially for output growth, and at times of heightened macroeconomic uncertainty. This result is consistent with a body of research which has emphasized the value of point forecasts from surveys of professional forecasters; e.g. see Krüger et al. (2017) and Clements (2018).

But we find that judgments about the variance (uncertainty) forecast tend to detract from the accuracy of short horizon density forecasts from statistical models, although they can help deliver better tail risk forecasts at long horizons. This mixed result on the utility of second-moment judgments is consistent with mixed evidence from previous research on the accuracy of the variance forecasts from the US SPF (Krüger et al. (2017) and Clements (2018)), the Euro SPF (Kenny et al. (2014; 2015)) and four central banks (Knüppel and Schultefrankenfeld (2019)).

Finally, judgments about skew do not in general improve density forecast accuracy. Kenny et al. (2014) also find skew forecasts from four central banks are not relevant. But skew judgments can improve tail risk forecasts for output growth at times of macroeconomic change, such as around business cycle turning points. This is because forecasts generated mechanically, even from flexible statistical models that generate asymmetric forecast densities, only adapt at a lag to changes in the state of the macroeconomy.

A timely observation, giving the emerging (at the time of writing) Covid-19 induced recession, is that at times of great uncertainty, our results do point to the value of judgmental-adjustments to macroeconomic forecasts from statistical models. Our results indicate that judgments about the central path of the economy have worked especially well in the aftermath of historical shocks. Future work will evaluate how well judgment-based macroeconomic forecasts performed relative to statistical models in the face of the Covid-19 pandemic.

References

Aastveit, K. A., Mitchell, J., Ravazzolo, F. and van Dijk, H. K. (2019). The evolution of forecast density combinations in economics, *Oxford Research Encyclopedia of Economics and Finance, Oxford University Press* .

Adrian, T., Boyarchenko, N. and Giannone, D. (2019). Vulnarable growth, *American Economic Review* **109**: 1263–89.

Altavilla, C., Giacomini, R. and Ragusa, G. (2016). Anchoring the yield curve using survey expectations, *Journal of Applied Econometrics* **32**: 1055–1068.

Britton, E., Fisher, P. and Whitley, J. (1998). The inflation report projections: understanding the fan chart, *Bank of England Quarterly Bulletin* **February**: 30–37.

Burgess, S., Fernandez-Corugedo, E., Groth, C., Harrison, R., Monti, F., Theodoridis, K. and Waldron, M. (2013). The Bank of England's forecasting platform: COMPASS, MAPS, EASE and the suite of models, *Bank of England Working Paper n. 471* .

Clark, T. E. (2011). Real-time density forecasts from bayesian vector autoregressions with stochastic volatility, *Journal of Business and Economic Statistics* **29**: 327–341.

Clements, M. P. (1995). Rationality and the role of judgment in macroeconomic forecasting, *The Economic Journal* **105**: 410–420.

Clements, M. P. (2014). US inflation expectations and heterogenous loss functions, 1968-2010, *Journal of Forecasting* **33**: 1–14.

Clements, M. P. (2018). Are macroeconomic density forecasts informative?, *International Journal of Forecasting* **34**(2): 181–198.

Clements, M. P. and Hendry, D. (1996). Intercept correction and structural change, *Journal of Applied Econometrics* **11**: 475–494.

Cogley, T., Morozov, S. and Sargent, T. (2005). Bayesian fan charts for U.K. inflation: Forecasting and sources of uncertainty in an evolving monetary system, *Journal of Economic Dynamics and Control* **29**: 1893–1925.

Davydenko, A. and Fildes, R. (2013). Measuring forecasting accuracy: the case of judgmental adjustments to SKU-level demand forecasts, *International Journal of Forecasting* **29**: 510–22.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of Diebold-Mariano tests, *Journal of Business and Economic Statistics* **33**: 1–24.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy, *Journal of Business and Economic Statistics* **13**: 253–263. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar.

Engelberg, J., Manski, C. F. and Williams, J. (2008). Comparing the point predictions and subjective probability distributions of professional forecasters, *Journal of Business and Economic Statistics* **27**: 30–41.

Fildes, R. and Stekler, H. (2002). The state of macroeconomic forecasting, *Journal of Macroeconomics* **24**(4): 435 – 468.

Giacomini, R. and Ragusa, G. (2014). Theory-coherent forecasting, *Journal of Econometrics* **182**: 145–155.

Giacomini, R. and Rossi, B. (2010). Forecast comparisons in unstable environments, *Journal of Applied Econometrics* **25**: 595–620.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability, *Econometrica* **74**: 1545 – 1578.

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules, *Journal of Business and Economic Statistics* **29**: 411–422.

Harvey, D. I., Leybourne, S. J. and Whitehouse, E. J. (2017). Forecast evaluation tests and negative long-run variance estimates in small samples, *International Journal of Forecasting* **33**: 833–847.

Hendry, D. and Clements, M. P. (2004). Pooling of forecasts, *Econometrics Journal* **7**: 1–31.

Hyndman, R. J. (2020). A brief history of forecasting competitions, *International Journal of Forecasting* **36**(1): 7 – 14.

Jore, A. S., Mitchell, J. and Vahey, S. P. (2010). Combining forecast densities from VARs with uncertain instabilities, *Journal of Applied Econometrics* **25**(4): 621–634.

Kenny, G., Kostka, T. and Masera, F. (2014). How Informative are the Subjective Density Forecasts of Macroeconomists?, *Journal of Forecasting* **33**(3): 163–185.

Kenny, G., Kostka, T. and Masera, F. (2015). Can Macroeconomists Forecast Risk? Event-Based Evidence from the Euro-Area SPF, *International Journal of Central Banking* **11**(4): 1–46.

Knüppel, M. and Schultefrankenfeld, G. (2012). How informative are central bank assessments of macroeconomic risks?, *International Journal of Central Banking* **8**(3): 87–139.

Knüppel, M. and Schultefrankenfeld, G. (2019). Assessing the uncertainty in central bank inflation outlooks, *International Journal of Forecasting* **35**(4): 1748–1769.

Krüger, F., Clark, T. E. and Ravazzolo, F. (2017). Using entropic tilting to combine BVAR forecasts with external nowcasts, *Journal of Business and Economic Statistics* **35**: 470–85.

Lawrence, M., Goodwin, P., O'Connor, M. and Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25years, *International Journal of Forecasting* **22**(3): 493 – 518.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T. (2017). Forecaster's dilemma: extreme events and forecast evaluation, *Statistical Science* **32**: 106–27.

McNees, S. K. (1990). The role of judgment in macroeconomic forecasting accuracy, *International Journal of Forecasting* **6**: 287–299.

Mitchell, J. (2005). The National Institute Density Forecasts of Inflation, *National Institute Economic Review* **193**: 60–69.

Pesaran, M. and Timmermann, A. (2007). Selection of estimation window in the presence of breaks, *Journal of Econometrics* **137**: 134–161.

Robertson, J. C., Tallman, E. W. and Whiteman, C. H. (2005). Forecasting using relative entropy, *Journal of Money, Credit and Banking* **37**(383-401).

Smets, F. and Wouters, R. (2007). Shocks and frictions in US business cycles., *American Economic Review* **97**: 586–606.

Tallman, E. W. and Zaman, S. (2020). Combining survey long-run forecasts and nowcasts with bvar forecasts using relative entropy, *International Journal of Forecasting* **36**(2): 373 – 398.

Timmermann, A. (2006). Forecast combinations, *in* G. Elliott, C. W. J. Granger and A. Timmermann (eds), *Handbook of Economic Forecasting Volume 1*, North-Holland, pp. 135–196.

Trapero, J. R., Pedregal, D. J., Fildes, R. and Kourentzes, N. (2013). Analysis of judgemental adjustments in the presence of promotions, *International Journal of Forecasting* **29**: 234–43.

Turner, D. S. (1990). The role of judgement in macroeconomic forecasting, *Journal of Forecasting* **9**(4): 315–345.

Table 1: Evaluating the effect of judgment on the WBS combination densities: "Tilting Approach 1"

Forecast Horizon	Statistical Model	Judgment: moments from professional forecasters							
h		Mean				Variance		Skew	
	WBS Comb.	MPC	NIESR	CE	SEF	MPC	SEF	MPC	SEF
Panel A: Output Growth (Forecast Origin: 2001Q1-2016Q1)									
	CRPS	Ratios between the CRPS of tilted and original density							
1	0.28	1.02	0.99	0.88		1.97		0.999	
4	0.93	0.85	0.86	0.79	0.88	1.00	1.00	0.997	1.022
8	1.09	1.03	**0.95**		**0.89**	0.99	1.01	0.999	1.019
	tw-CRPS	Ratios between the *tw*-CRPS of tilted and original density							
1	0.07	0.68	0.99	0.87		3.03		1.002	
4	0.16	1.27	0.98	1.11	0.78	0.88	**0.83**	0.982	1.060
8	0.09	1.40	1.04		0.79	0.90	**0.30**	**0.950**	1.042
Panel B: Inflation (Forecast Origin: 2004Q1-2016Q1)									
	CRPS	Ratios between the CRPS of tilted and original density							
1	0.14	**0.79**	1.59	0.80		1.70		1.004	
4	0.68	0.87	1.09	0.86	1.05	1.00	1.04	1.002	1.004
8	0.86	0.97	0.94		0.88	**0.98**	1.08	1.006	1.007
	tw-CRPS	Ratios between the *tw*-CRPS of tilted and original density							
1	0.04	0.72	1.06	0.44		1.98		1.001	
4	0.17	0.75	1.05	0.49	0.49	1.03	0.92	**0.986**	0.996
8	0.20	0.93	**0.75**		**0.61**	1.02	**0.70**	0.974	0.939
Panel C: Output Growth and Inflation (across forecast horizons and loss functions)									
Prop. Improved		72%	69%	86%	88%	38%	50%	34%	13%
Prop. Sign. Improved		3%	9%	16%	25%	6%	38%	6%	0%
Prop. Worsened		9%	3%	0%	0%	13%	25%	6%	13%

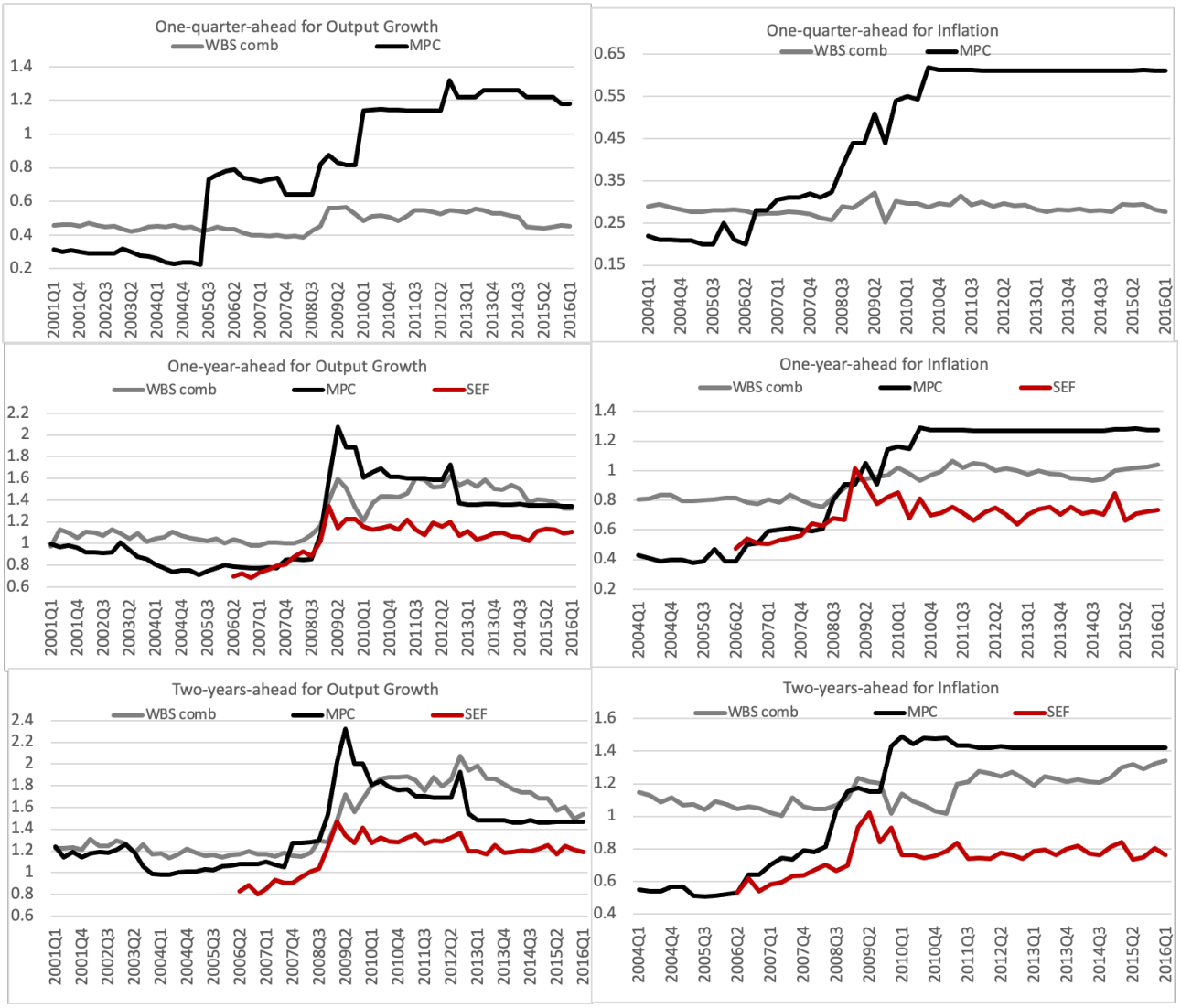
Notes: Tilting Approach 1 imposes one judgment-based predictive moment at a time, keeping the three other moments (up to the 4th moment) at the statistical model values. Values in bold: statistically significant improvement in forecast accuracy due to judgment. Underlined: statistically significant worsening due to judgment. These are based on a two-sided 10% level test of the null of equal forecast accuracy with small sample correction, see eq. (14). The *tw*-CRPS thresholds are $r_1 = 0$ and $r_2 = 4\%$, and these are for year-on-year growth rates. Results for SEF for forecast origins from 2006Q2 onwards only. Panel C: *Prop. Improved* denotes the percentage of occasions (across variables (output growth and inflation, forecast horizons (h=1 to 8) and scores (CRPS and tw-CRPS)) that judgment improves the accuracy of the statistical model's density forecast. *Prop. Sign. Improved* denotes the percentage of occasions that this improvement is statistically significant at the 90% significance level (using individual *t*-tests). *Proportion Sign. Worsened* denotes the percentage of occasions that judgment leads to a statistically significant worsening of forecast performance.

Table 2: Evaluating the effect of judgment on the WBS combination densities: "Tilting Approach 2"

Forecast Horizon	Statistical Model	Judgment: moments from professional forecasters									
h		Mean		Variance		Skew		Mean + Var		Mean + Var + Skew	
	WBS Comb.	MPC	SEF	MPC	SEF	MPC	SEF	MPC	SEF	MPC	SEF
Panel A: Output Growth (Forecast Origin: 2001Q1-2016Q1)											
	CRPS	Ratios between the CRPS of tilted and original density									
1	0.28	1.04		1.29		1.03		1.37		1.61	
4	0.93	0.84	0.88	1.01	1.01	1.02	1.02	0.84	0.87	0.83	0.87
8	1.09	1.05	**0.88**	1.01	1.04	1.03	0.99	1.05	**0.88**	1.03	**0.89**
	tw-CRPS	Ratios between the *tw*-CRPS of tilted and original density									
1	0.07	0.65		1.60		1.04		1.28		1.83	
4	0.16	1.00	**0.60**	0.90	**0.77**	**0.76**	**0.67**	1.12	0.62	1.25	0.66
8	0.09	**0.39**	**0.22**	0.88	**0.41**	**0.30**	**0.37**	1.15	**0.23**	1.31	**0.28**
Panel B: Inflation (Forecast Origin: 2004Q1-2016Q1)											
	CRPS	Ratios between the CRPS of tilted and original density									
1	0.14	**0.70**		1.22		0.99		1.37		1.41	
4	0.68	0.91	1.14	1.03	1.02	1.02	1.03	0.84	1.12	0.90	1.11
8	0.86	1.04	0.97	1.03	1.11	1.01	1.02	1.05	0.97	0.99	0.97
	tw-CRPS	Ratios between the *tw*-CRPS of tilted and original density									
1	0.04	0.46		1.22		1.02		1.28		1.38	
4	0.17	0.61	**0.27**	1.14	**0.91**	0.94	0.92	1.12	**0.30**	0.85	0.31
8	0.20	0.77	**0.28**	1.21	1.73	0.78	0.81	1.15	**0.23**	1.24	**0.24**
Panel C: Output Growth and Inflation (across forecast horizons and loss functions)											
Prop. Improved		75%	88%	16%	38%	38%	63%	38%	88%	38%	88%
Prop. Sign. Improved		13%	63%	0%	38%	25%	6%	0%	50%	0%	38%
Prop. Sign. Worsened		0%	0%	25%	25%	0%	0%	6%	0%	16%	0%

Notes: See notes to Table 1. Unlike Tilting Approach 1, Tilting Approach 2 does not constrain the other three moments (up to the 4th moment) at the statistical model values.

Figure 1: Predicted uncertainty (standard deviation) for MPC, SEF and WBS combination forecasts



Notes: The horizontal axis dates refer to the forecast origin. The MPC's predicted standard deviation is computed from the parameters of the two-piece normal density. SEF standard deviation is computed from a generalized beta distribution fitted to the aggregated SEF histograms.

Figure 2: Predicted skew by professional forecaster

Figure 2A: MPC forecasts one, four and eight-quarters-ahead

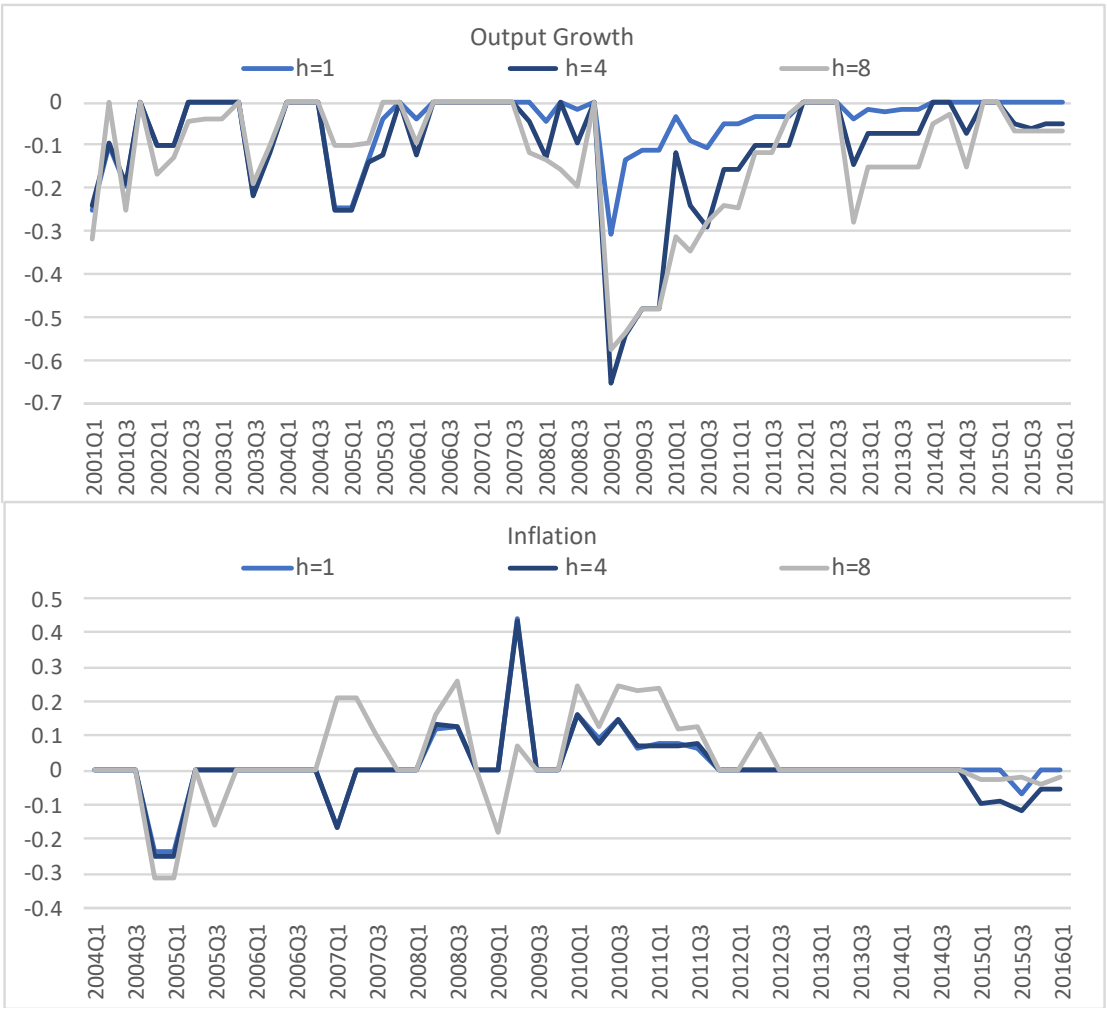
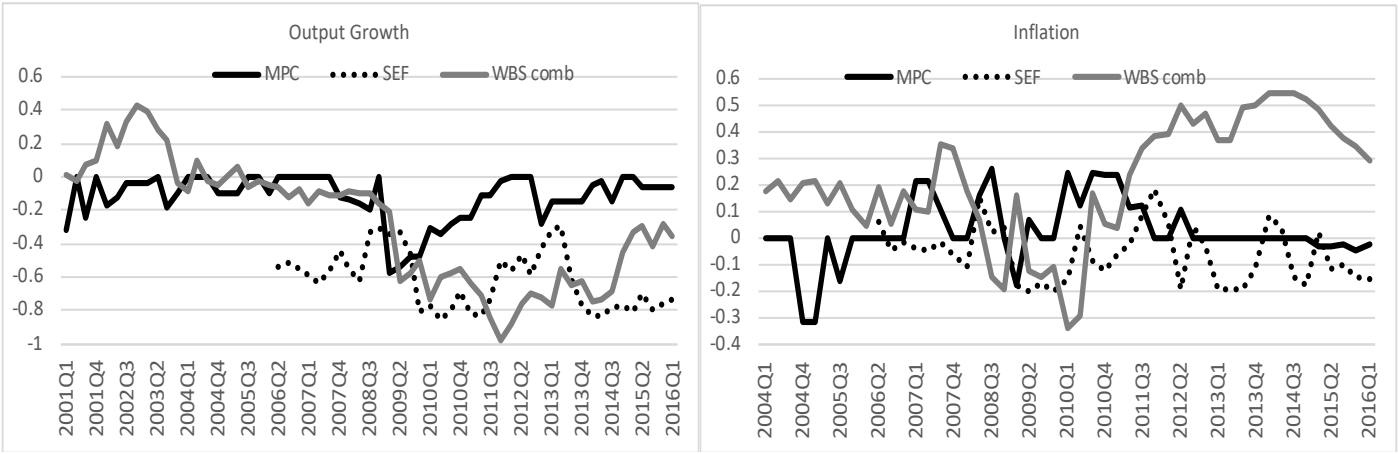


Figure 2B: MPC, SEF and WBS combination predicted skew for the two-year-ahead forecasts



Notes: The x-axis dates refer to the forecast origin. The MPC's predicted skew is computed from the parameters of the two-piece normal density. SEF skew is computed from a generalized beta distribution fitted to the aggregated SEF histograms.

Figure 3: The time-varying effects of judgment: t-statistics from the fluctuation test, computed against the WBS combination density, over a rolling window of 20 quarters using CRPS and tw-CRPS loss

Figure 3.1: MPC and CE predicted mean judgments for output growth one-year-ahead

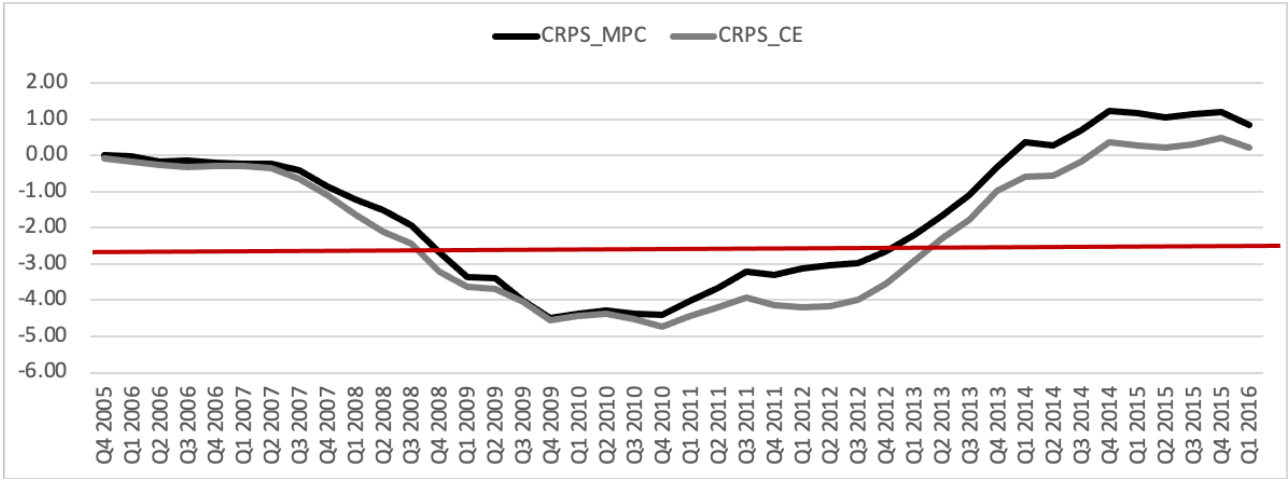


Figure 3.2: MPC predicted variance judgments for output growth and inflation one-quarter-ahead

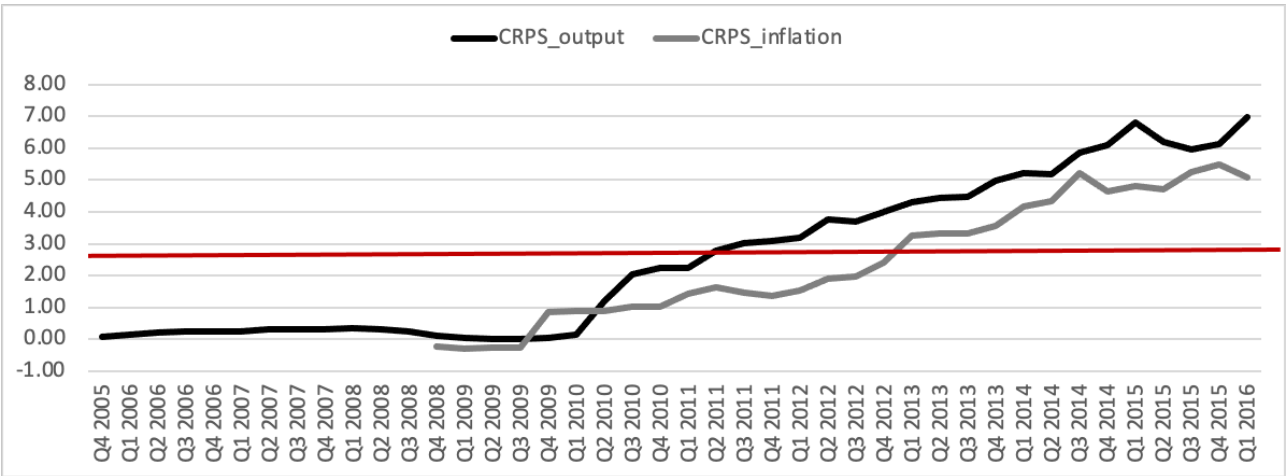


Figure 3.3: MPC predicted uncertainty judgments for output growth and inflation two-years-ahead

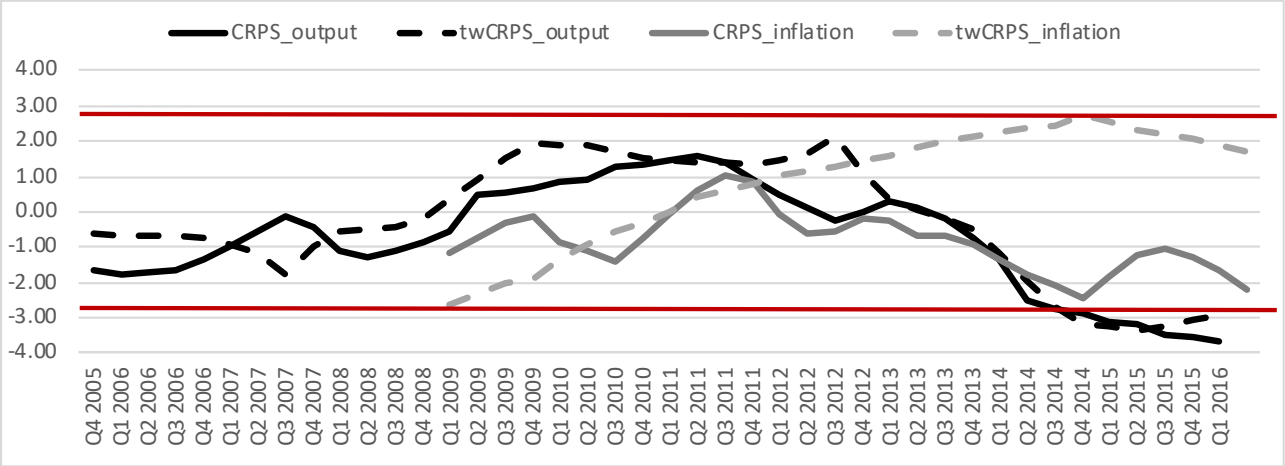
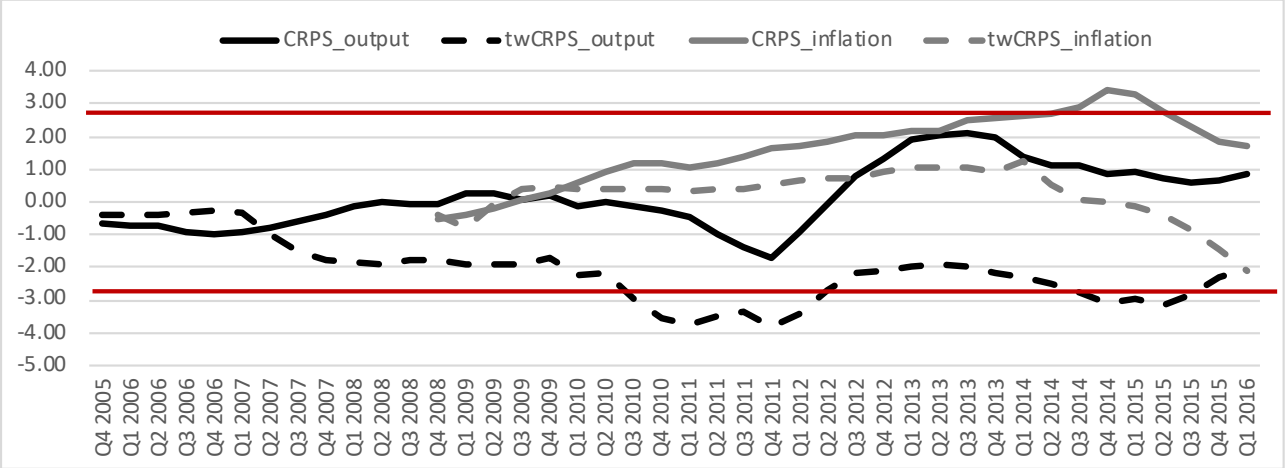


Figure 3.4: MPC predicted skew judgments for output growth and inflation two-years-ahead



Notes: The dates reported refer to the last forecast origin, τ, included in the rolling window. The horizontal lines are the critical values for the fluctuation test. The two-sided 10% test critical values are 2.8 for output growth and 2.5 for inflation, but the values indicated are for output growth. Values above the positive critical value indicate that judgment worsens the statistical model's forecasting performance. Values below the negative critical value indicate that judgment improves the statistical model's forecasting performance.

Online Appendix for: "Does Judgment Improve Macroeconomic Density Forecasts?"

Ana Beatriz Galvao
University of Warwick

Anthony Garratt
University of Warwick

James Mitchell
Federal Reserve Bank of Cleveland

January 2021

A Professional Forecasters

A.1 National Institute of Economic and Social Research (NIESR)

The NIESR produce quarterly macroeconomic forecasts using their Global Economic Model (NiGEM). This model is also used by many European central banks and international organizations, such as the OECD.[1] When undertaking forecasting, and conducting policy-based analysis, residual adjustments are implemented, judgmental assumptions are made and paths are set for a range of exogenous variables. As such, and as Turner (1990) concludes in an earlier analysis of the role of judgment on NIESR point forecasts, it is not possible for the typical forecast user to isolate the direct quantitative role of judgment on NIESR's forecasts even when re-running NiGEM under alternative judgmental adjustments.

A.2 The Bank of England's Monetary Policy Committee

The MPC's density forecasts for GDP growth and inflation are communicated as fan charts in the Bank of England's quarterly *Inflation Report*. The fan charts, as stated by the MPC: "represent the MPC's best collective judgment about the most likely paths for inflation and output, and the uncertainties surrounding those central projections". Since the establishment of the MPC in 1997 the fan charts have been constructed using two-piece normal densities. The

[1] NIESR, founded in 1938, has a long track-record of producing quarterly macroeconomic forecasts. NiGEM is an estimated macro-econometric model for a group of countries based on a New-Keynesian framework, with agents assumed forward-looking and nominal rigidities slowing the adjustment process to external shocks or events. The model pays particular attention to its long-term equilibrium properties; while the short-term dynamics and underlying estimated properties of the model are consistent with data and result from well documented and robust estimation methods. See https://nimodel.niesr.ac.uk/index.php?t=5NiGEM

two-piece normal density, $f^{2PN}(Y)$, creates a skewed density by combining the two halves of two normal densities with a common mode of μ and standard deviations of σ_1 and σ_2:

$$f^{2PN}(Y) = \begin{cases} A \exp\left[-(y-\mu)^2/2\sigma_1^2\right] & \text{if } y < \mu \\ A \exp\left[-(y-\mu)^2/2\sigma_2^2\right] & \text{if } y \geq \mu \end{cases} \tag{1}$$

where $A = \left(\sqrt{2\pi}\left(\sigma_1 + \sigma_2\right)/2\right)^{-1}$. $f^{2PN}(Y)$ has computational advantages over other skewed distributions given it can be analyzed using analytical formula; see Wallis (2014) for a historical perspective. The two-piece normal can be parameterized in alternative ways to (1). Given this, as discussed by Wallis (2004), care needs to be exercised when using the published parameters of the two-piece normal to back out estimates of μ, σ_1 and σ_2, as the Bank prefer an alternative parameterization. The two-piece normal is asymmetric when $\sigma_1 \neq \sigma_2$.

The MPC choose to communicate skewness by emphasizing mode skewness: $(E(Y) - \mu)$. But in our main paper we use the Pearson moment coefficient of skewness i.e. the third standardized moment of Y, as an input into the entropic tilting. The MPC forecasts are conditional on forward market interest rates and other conditioning assumptions detailed in the *Inflation Report,* including about the stock of purchased assets.

A.3 The Bank of England's Survey of External Forecasters

The Survey of External Forecasters (SEF) has run quarterly since 1996 and asks a panel, of typically twenty to thirty professional forecasters, for their probabilities that the future value of the variable of interest will lie within a number of pre-assigned intervals. Thereby individual-level forecast histograms are defined. The forecasters surveyed include City of London firms, academic institutions and private consultancies mainly based in London (see Boero, Smith and Wallis (2008) for further details). Originally, the SEF asked for point and density forecasts for just inflation and on a "fixed-target" basis, meaning the forecast horizon changed from one survey to the next.

A.4 Consensus Economics

Consensus Economics (CE), founded in 1989, is a leading international economic survey organisation that polls economists to obtain their latest forecasts. In the UK, around 30-40 professional forecasters provide their forecasts to CE. Forecasters are asked for their point forecasts only.

Consensus Economics do provide, at some forecast horizons, supplementary information on *disagreement* i.e. the degree to which forecasters disagree with the average forecast. As we aim to quantify the effects of forecast uncertainty judgments, we chose not to use *disagreement* as a proxy for forecast uncertainty.

B The Statistical Models

In this appendix, we describe the estimation and forecasting details of the *three* statistical forecasting models considered in the main paper: (i) the WBS combination; (ii) the BVAR model; and (iii) the AR model.

To mimic real-time application, each of the three statistical models is re-estimated at each forecast origin, τ, over the out-of-sample window, denoted $\tau = T+1, ..., T+P$, using an expanding window of data. Denote the h-quarter ahead density forecasts of the random variable $Y_{\tau+h}$, made in the middle of quarter τ, $f_{\tau,h}(Y)$.

B.1 The WBS Combination

The WBS combined density forecasts are produced by combining the density forecasts across the 24 statistical forecasting models described below. Each forecasting model is classified within three *classes* of model.

B.1.1 Class 1: AR models

We employ an AR(2) model, estimated using quarterly growth rate data defined as $y_t = 400 \times (\ln(z_t) - \ln(z_{t-1}))$, where z_t is either the level of real GDP or the Consumer Price Index ($t = 1, 2, ..., \tau$).[2] Given the parameters estimated using data up to the forecast origin, τ, we use the model to generate sequences of $h = 1, ..., 8$ quarter ahead forecast draws. These draws are obtained through iteration using the bootstrap methods described in Clements and Taylor (2001), taking into account both parameter and forecast uncertainty. This method involves first bootstrapping from the residuals and uses the estimated autoregressive parameters to obtain a full bootstrapped time-series for y_t^* ($t = 3, ..., \tau$). Then the AR(2) model is re-estimated, for each of the bootstrapped samples, with h-quarter ahead forecasts computed by iteration

[2] The choice of an AR(2) specification is a fairly popular benchmark (Carriero, Galvao and Kapetanios, 2019). There is evidence, particularly for output growth, that an AR(2) model forecasts competitively; e.g. see Chauvet and Potter (2013).

while bootstrapping from the new set of residuals to accommodate forecast uncertainty. We use the sequence of $h = 1, ..., 8$-quarter ahead forecasts for the quarterly growth rates to compute forecasts of year-on-year growth rates $(100 \times ((z_t/z_{t-4}) - 1))$, given known values for z_τ and its lags. In our empirical application, we use $5,000$ bootstrap replications for y_t^* to generate $5,000$ draws from the predictive density, $f_{\tau,h}(Y)$, for each horizon, h.

B.1.2 Class 2: BVAR models

Three types of BVAR are considered and differ according to the (quarterly) variables considered and whether stochastic volatility is modeled. In total, the BVAR model class comprises 10 models.

Firstly, a single macro BVAR *without* stochastic volatility models the log-levels of the seven variables used in the Smets and Wouters (2007) DSGE model: GDP, CPI, consumption, investment, hours, real wages and the interest rate (the Bank Rate is used). Details of the data sources and data transformations used are provided in Table A6. The lag order p is set to four. We generate posterior densities of the autoregressive coefficients using a Minnesota prior. The overall tightness around these autoregressive parameters is controlled via priors on the hyperparameters, which are set to maximize the marginal data density as in Carriero, Clark and Marcellino (2015) and Carriero, Galvao and Kapetanios (2019).

The second type of BVAR model, is a set of medium-sized BVARs *without* stochastic volatility. In addition to GDP and CPI inflation, these VAR models include the thirteen indicator variables as described in Table A6 - in the 'medium-size models' panel. We estimate eight variants of this medium-sized BVAR, for lag lengths $p = 1, \ldots, 4$, in both log-levels and first-differences of the data.

Finally, we consider a single medium-sized BVAR *with* stochastic volatility in these same 15 variables. There is evidence that accommodating stochastic volatility is especially helpful when density forecasting; e.g. see Clark (2011). The variables are modeled in first differences and the lag order p is set to four. Estimation follows Carriero, Clark and Marcellino (2019). These methods allow for the estimation of large BVARs with non-conjugate priors and drifting volatilities based on a triangularization of the system. This reduces the computational burden. It uses draws from a BVAR with an independent Normal Wishart prior and stochastic volatility. The predicted densities are obtained using a set of kept posterior density draws.

B.1.3 Class 3: Mixed-Frequency models

The second class of model comprises mixed-frequency models. These allow for the consideration of known (within-quarter) monthly information reflecting the release calendars of a set of monthly indicators. Specifically, Autoregressive Distributed Lag Mixed Data Sampling (ADL-MIDAS) models are estimated for each horizon and target variable (quarterly GDP growth and quarterly CPI inflation). The models are single predictor ADL-MIDAS regressions, using in turn each of the thirteen indicators included in the medium-sized BVAR (see Table A6), measured at a monthly frequency. The ADL-MIDAS models are estimated by nonlinear least squares assuming a beta-weighting function. Estimation involves single predictor models with autoregressive terms, predictors are sampled monthly, autoregressive terms are quarterly for output growth, monthly for inflation, and the number of monthly lags used is 24. We use the values for the first month of the current quarter for a subset of predictors (financial variables) that are published rapidly; and lagged values of inflation. The data transformation (log or log-differences) differs across variables and is described in Table A6. The predicted densities for GDP growth and inflation are computed using a fixed-regressor bootstrap approach, as described in Carriero, Galvao and Kapetanios (2019), taking into account both parameter and forecast uncertainty.

B.1.4 Combining the predictive densities

Based on the methods described above, we take $4,000$ draws from the predictive density of each of the 24 statistical forecasting models. A nonparametric kernel density estimator is fitted to these draws, to obtain a continuous predictive density over a fixed grid of $1,000$ values lying between -15 and 15 annual percent, for both output growth and inflation.

We then combine these out-of-sample density forecasts from the $N = 24$ forecasting models using the Logarithmic Opinion Pool (LogOP). Given $i = 1, \ldots, N$ forecasting models, the LogOP combined output growth or inflation density forecast made at time τ for $y_{\tau+h}$ is:

$$f_{\tau,h}(Y) = \frac{\prod_{i=1}^{N} g_{\tau,h}(Y \mid I_{i,\tau})^{w_{i,\tau}}}{\int \prod_{i=1}^{N} g_{\tau,h}(Y \mid I_{i,\tau})^{w_{i,\tau}} dY}, \tag{2}$$

where $g_{\tau,h}(y_\tau \mid I_{i,\tau})$ is the $h-$step-ahead output growth or inflation density forecast from model i, conditional on the information set $I_{i,\tau}$. The non-negative weights, $w_{i,\tau}$, sum to unity, where

the weights change with each recursion in the evaluation period, τ. The denominator in (2) is a constant that ensures that the combined density is a proper density.

The weights, $w_{i,\tau}$, in (2) are estimated using the logarithmic score. We note, however, that other loss functions such as the CRPS could also be used. The intuitive appeal of the logarithmic scoring rule stems from the high score it assigns to a density forecast with high probability at the subsequently realized value. The logarithmic score of the i^{th} density forecast, $\ln g_{\tau,h}(y_{\tau+h} \mid I_{i,\tau})$, is the logarithm of the density forecast $g_{\tau,h}(. \mid I_{i,\tau})$, evaluated at the realization, $y_{\tau+h}$.

Specifically, the recursive weights are updated as follows:

$$w_{i,\tau} = \frac{\exp\left[\sum_{\tau^*=\tau-h-12}^{\tau-h} \ln g_{\tau^*,h}(y_{\tau^*+h} \mid I_{i,\tau^*})\right]}{\sum_{i=1}^{N} \exp\left[\sum_{\tau^*=\tau-h-12}^{\tau-h} \ln g_{\tau^*,h}(y_{\tau^*+h} \mid I_{i,\tau^*})\right]} \tag{3}$$

It is important to note that the weights on the various component densities in (3) vary through time, τ, reflecting the rolling historical performance of the component density forecasts. Hence, the combination potentially exhibits greater flexibility than any single component density forecast (in which the individual model parameters are themselves recursively updated). The logarithmic score is computed as an average over a rolling window of the last 12 quarterly observations. This allows the weights in the combination to adapt quickly to changes in relative model performance. Only forecasting errors actually known in real-time feed into how these weights are calculated. Using these weights, a combined density forecast is then produced across the 24 models.

As reviewed by Aastveit, Mitchell, Ravazzolo and van Dijk (2019), the linear opinion pool remains the most popular means of producing combined densities in the macroeconomic forecasting literature. However, when implementing the WBS Forecasting System, in 2014 the decision (based on preliminary evaluation results at the time) was taken to use the logarithmic opinion pool.[3] It was found to produce more accurate density forecasts for UK GDP growth and inflation than the linear opinion pool. Therefore we stick with the logarithmic pool in this paper. Favoring the logarithmic pool is consistent with results showing that the linear opinion pool can overstate forecast uncertainty; see Knüppel and Krüger (2019). The logarithmic pool is known to preserve the distributional form of the component densities when they are all from the same exponential family (e.g. see Wallis (2011)). Given that many of our component

[3] Since 2014 these forecasts have been published at https://warwick.ac.uk/fac/soc/wbs/subjects/finance/mpf/forecasting/

densities are computed via bootstrap, which does not require the assumption that disturbances are Gaussian, the logarithmically combined density forecast can still generate non-Gaussian asymmetric densities.

B.2 The BVAR Model

We consider a small BVAR model with five endogenous variables, all in growth rates. The variables included in the VAR, in addition to GDP and CPI, are the unemployment rate, the three-month interest rate and the real effective exchange rate index - to reflect the fact that the UK is a small open economy.

We estimate the BVAR with a Minnesota prior, with prior hyper-parameters set to maximize the marginal data density as described in Carriero et al. (2015) and Carriero, Galvao and Kapetanios (2019) with the autoregressive lag order set to 4. We use draws from the normal posterior distribution of the dynamic coefficients and from the Wishart posterior distribution of the variance-covariance matrix to obtain sequences of density forecast draws, by iteration, for $h = 1$ to 8 quarters ahead. As well as computing the sequence of forecasts for a given draw from the parameters, we also accommodate forecasting uncertainty by drawing from the errors. The number of draws from the predictive distribution for each horizon is set to $5,000$. As before, we transform the quarterly growth rate forecasts to obtain forecasts for year-on-year growth rates.

B.3 The AR Model

We estimate an AR(2) model by OLS. Details of how the predictive density are computed for each variable and forecast horizon are in section B.1.1.

C Supplementary Empirical Results and Computational Details

This appendix contains a series of additional empirical results referred to in the main paper.

C.1 Robustness of densities fitted to SEF histogram forecasts

As a robustness check, we calculate the mean and variance of the SEF histograms nonparametrically as described in Clements (2019) and D'Amico and Orphannides (2008). This assumes that the probability mass is concentrated at the mid-point of each bin. These nonparametric

results are reported in Figures A1-A4. For both inflation and output growth, the nonparametric mean estimates are very close to those calculated fitting normal and generalized beta densities to the histograms. The variances for inflation tend to be, on average, slightly larger, but with very strong comovement. This was also documented, by Clements (2019), when making the same the variance comparison for the US Survey of Professional Forecasters' assessments of inflation. For output growth, the variances calculated directly from the histograms are close to those computed using the normal and generalized beta densities. Given these similarities and the advantages cited in Engelberg, Manski and Williams (2008)) of imposing distributional assumption enabling "sharper empirical analysis", we proceed in the main paper using the generalized beta, as it allows for asymmetric densities.

C.2 Comparison of uncertainty forecasts against ex post estimates

Here we compare the uncertainty forecasts from the professional forecasters, referred to by Clements (2014) as *ex ante* forecasts of uncertainty, with *ex post* estimates computed as the standard deviation of the historical mean squared point forecast error. See also Jo and Sekkel (2019) and Clark, McCracken and Mertens (2020). This analysis serves as a test of whether judgment about the second moment improves upon an unconditional estimate that assumes the average level of uncertainty that has been experienced in the recent past will continue into the future. Clements (2018) undertakes a similar comparison for the US Survey of Professional Forecasters, finding that second-moment judgment makes simple statistical density forecasts worse. Knüppel and Schultefrankenfeld (2019) finds no evidence for their bias, but like Clements (2018) finds four central banks' judgment-based assessments of inflation uncertainty to be underconfident at short horizons.

Since the CE, NIESR and SEF point forecasts are similar to those from the MPC, we focus here on constructing these ex post estimates of forecast uncertainty using the MPC's point forecasts. We further focus on the MPC's output growth forecasts. Recall that the sample of historical point forecasts from the MPC is shorter for inflation.

We calculate the standard deviation over a rolling window of the last 16 point forecasting errors made by the MPC. We use a rolling window to allow the standard deviation to change over time, reflecting changes in historical forecasting performance. We calculate estimates as if in real-time; i.e. we take into account that when producing a forecast at τ, we can only consult forecasting errors up to $\tau - h$. We use the second monthly release of GDP as the realization

against which the forecast is compared. An implication of this is that at long horizon forecasts the ex post predictive standard deviation will only adjust at a lag to accommodate recent changes in underlying forecasting accuracy.

Comparison in Figure A5 of the ex post uncertainty estimates with the ex ante estimates, for the MPC's output growth forecasts, reveals that movements in ex post forecast uncertainty are not matched by the ex ante forecasts. This indicates that the MPC's assessments of uncertainty do not directly reflect historical point forecasting performance. For the one-quarter-ahead forecasts, the MPC's ex ante estimates exceed the ex post estimates - suggesting that the MPC overstated the uncertainty in their forecasts. Clements (2014; 2018) and Knüppel and Schultefrankenfeld (2019) found similar results (for the US SPF and four central bank forecasts, including the Bank of England). They discuss the tendency for uncertainty forecasts to be underconfident at short horizons. For correctly calibrated uncertainty forecasts, in population, we should expect equality of the ex ante and ex post forecasts as discussed in Clements (2014). But at the longer horizons (bottom two panels of Figure A5), while less volatile, we do see the MPC's ex ante uncertainty forecasts move more in-line with changes in their recent forecast performance, as measured by the ex post forecast standard deviation. This is especially so around the time of the financial crisis, in 2008-2009. We also observe a tendency for the MPC to perceive more uncertainty than the ex post estimates during relatively tranquil periods (such as 2004 and 2016). In contrast, during more turbulent periods (like the aftermath of the financial crisis), their assessment of the forecast uncertainty is lower than the ex post uncertainty forecasts. These differences confirm the judgment component to the MPC's published (i.e. ex ante) uncertainty forecasts.

C.3 Computation of tilted densities

As described in the main paper, the tilted density $\widetilde{f}_{\tau,h}(Y)$ is computed as:

$$\widetilde{f}_{\tau,h}(Y) = f_{\tau,h}(Y) \exp\left\{\eta_{\tau,h} + \tau'_{\tau,h}(g_{\tau,h}(Y) - \bar{g})\right\}, \tag{4}$$

where $\eta_{\tau,h}$ and $\tau_{\tau,h}$ are obtained by numerical approximation of the underlying integrals:

$$\tau_{\tau,h} = \arg\min_{\tau} \frac{1}{D} \sum_{d=1}^{D} f_{\tau,h}(y^d) \exp\left\{\tau'(g_{\tau,h}(y^d) - \bar{g})\right\} \tag{5}$$

$$\eta_{\tau,h} = \log\left\{\frac{1}{D} \sum_{d=1}^{D} f_{\tau,h}(y^d) \exp[\tau'_{\tau,h}(g_{\tau,h}(y^d) - \bar{g})]\right\}^{-1}, \tag{6}$$

where $\{y^d\}_{d=1}^{D}$ are draws from $f_{\tau,h}(Y)$. We set $D = 10,000$.

For the BVAR model, we follow Robertson, Tallman and Whiteman (2005) and set the probability $f_{\tau,h}(y^d) = 1/D$ for each draw obtained via simulation. For the WBS combination, the predictive density is a combination of densities that have been estimated by applying a nonparametric density estimator to the component density draws. Using this nonparametric density, D random draws, y^d, are then obtained. We apply the Matlab function *randsample()* using as input the nonparametric probabilities computed for each value in a grid of $1,000$ points. The probabilities for each draw, i.e. $f_{\tau,h}(y^d)$, are computed by approximating the density with a mixture of two-normals such that the probabilities are computed using a closed-form solution. We use the Matlab procedure *fitgmdist* that employs the EM algorithm to estimate the parameters of the mixture of normals using the draws y^d for $d = 1, ..., D$.

C.4 Computation of CRPS

The CRPS evaluates the 'whole' density, while the threshold-weighted CRPS focuses on accuracy in the tails.

As explained in the main paper, the CRPS is given as:

$$CRPS_{t,h} = \int_{-\infty}^{+\infty} \left[F_{t,h}(y) - I(y_{t+h} \leq y)\right]^2 dy \tag{7}$$

where $F_{t,h}(.)$ is the CDF associated with the density forecast $f_{t,h}(.)$ and $I(y_{t+h} \leq y)$ denotes an indicator function equal to unity if $y_{t+h} \leq y$, 0 otherwise. As do not know the parametric form of the predictive densities from the statistical models, we compute the CRPS using the empirical CDF via equation (9) in Krüger, Lerch, Thorarinsdottir and Gneiting (2020).

The threshold-weighted CRPS is:

$$twCRPS_{t,h} = \int_{-\infty}^{+\infty} w(y) \left[F_{t,h}(y) - I(y_{t+h} \leq y)\right]^2 dy \tag{8}$$

where $w(y)$ are positive weights. $twCRPS_{t,h}$ is a *proper* scoring function; cf. Lerch, Thorarinsdottir, Ravazzolo and Gneiting (2017). When $w(y) \equiv 1$ for all y, $twCRPS_{t,h}$ reduces to the unweighted CRPS, (7). Otherwise, $w(y)$ can be tailored to focus on specific regions of the density. We specify $w(y)$ to focus on the tails. If we define a Gaussian CDF for y as $\Phi(y, mean, var)$ and let the two thresholds, r_1 and r_2, define regions in the left and right tails, then $w(y)$ is set

as:

$$w(y) = (1 - \Phi(y, r_1, var)) + \Phi(y, r_2, var). \tag{9}$$

Computationally, we compute $twCRPS_{t,h}$ via draws y^d from the predictive density. We first sort the draws in increasing order, then we calculate:

$$twCRPS_{t,h} = \frac{2}{D^2} \sum_{d=1}^{D} [(y^d - y_{t+h})(DI(y_{t+h} < y^d) - d + 0.5)] w(y^d),$$

where $w(y^d)$ applies the weight function described in the main paper for each draw, $I()$ is an indicator function and D is the total number of draws. The formula above is a small modification of equation (9) in Krüger et al. (2020).

D Examples of Tilted Densities using "Tilting Approach 1"

While the literature provides numerous examples illustrating the effects of exponentially tilting towards a judgment-augmented point forecast (e.g. see Altavilla, Giacomini and Ragusa (2016)), there are fewer studies that focus on tilting towards judgment-augmented uncertainty and skew forecasts. In this appendix, we accordingly illustrate how Tilting Approach 1 (as described in the main paper) affects the shape of the forecast densities. We do so by tilting the output growth and inflation forecasts from the WBS combination at $h = 1, 8$ towards the MPC's judgments on uncertainty and skew (second and third moments).

We choose a specific forecast origin – 2009Q2 – at the height of the Great Recession to illustrate the effects of tilting on the shape of the densities. From Figure 1 (in the main paper), we see that at this forecast origin the MPC's predicted standard deviation forecast for output growth is substantially higher than implied by the WBS combination at both $h = 1, 8$. This is because, in the face of the emerging recession, the MPC rapidly increased its explicit standard deviation forecasts during 2009. As Figure 2 (in the main paper) shows, this is accompanied by the MPC forecasting negative skew, especially at $h = 8$: the MPC's skew forecast is -0.53. But the WBS combination also predicts this, similarly forecasting negative skew of -0.60. For inflation, the MPC forecast more uncertainty than the WBS combination at $h = 1$ although at $h = 8$ their forecasts are more similar. The skew forecasts for two-years-ahead inflation from the WBS combination are small and negative, but small and positive for the MPC. For the one-quarter-ahead forecasts, both skew forecasts are positive, but the MPC's predicted skew

(0.45) is larger than the WBS combination (0.09).

Figure A7 compares the original (WBS combination) forecast, $f_{2009Q2,h}(Y)$, with the tilted densities, $\tilde{f}_{2009Q2,h}(Y)$. We illustrate the effects of tilting both towards the MPC's uncertainty and skew forecasts. Each figure includes a bar to represent the subsequent realization, y_{t+h}.

For the one-quarter-ahead forecasts, the left-hand-side panels of Figure A7 show that tilting the fairly bell-shaped WBS combination densities towards the higher MPC's predicted variance increases the tail probabilities in both of the tilted densities. The tilted densities exhibit local peaks (multi-modalities) in both tails. We note that if one did not constrain the other moments of the tilted density to remain the same as the WBS density, tilting towards the higher variance delivers a fairly similar shaped density but with changes to the density skewness and kurtosis; see Figures A8 and A9 in the online appendix.

Turning to the two-year-ahead forecasts, for inflation (bottom right panel of Figure A7) we see that the tilted density is similar to the original density. This is to be expected when the moments of the original density are fairly close to the variance and skew forecasts from the MPC. But for output growth (top right panel of Figure A7), we see changes relative to the original density when tilting towards the MPC's variance forecast (but not its skew forecast, as this is similar to the original skew forecast). Targeting the higher uncertainty forecast by the MPC, we see that the tilted density exhibits more probability mass in the tails, especially the right-tail where a peak is observed. Again we note that if one did not constrain the other moments of the tilted density to remain the same as the WBS density, tilting towards the higher variance delivers a similarly shaped density but with changes to the density skewness and kurtosis; see Figure A10.

References

Aastveit, K. A., Mitchell, J., Ravazzolo, F. and van Dijk, H. K. (2019). The evolution of forecast density combinations in economics, *Oxford Research Encyclopedia of Economics and Finance, Oxford University Press* .

Altavilla, C., Giacomini, R. and Ragusa, G. (2016). Anchoring the yield curve using survey expectations, *Journal of Applied Econometrics* **32**: 1055–1068.

Boero, G., Smith, J. and Wallis, K. F. (2008). Uncertainty and disagreement in economic prediction: the Bank of England survey of external forecasters, *Economic Journal* **118**: 1107–

1127.

Carriero, A., Clark, T. E. and Marcellino, M. (2015). Bayesian VARs: Specifications choices and forecast accuracy, *Journal of Applied Econometrics* **30**: 46–73.

Carriero, A., Clark, T. E. and Marcellino, M. (2019). Large vector autoregressions with stochastic volatility and non-conjugate priors, *Journal of Econometrics* **212**: 137–54.

Carriero, A., Galvao, A. and Kapetanios, G. (2019). A comprehensive evaluation of macroeconomic forecasting methods, *International Journal of Forecasting* **35**: 1226–1239.

Chauvet, M. and Potter, S. (2013). Forecasting output, *Handbook of Economic Forecasting, volume 2A*, Elsevier, chapter 3, pp. 141–194.

Clark, T. E. (2011). Real-time density forecasts from bayesian vector autoregressions with stochastic volatility, *Journal of Business and Economic Statistics* **29**: 327–341.

Clark, T. E., McCracken, M. W. and Mertens, E. (2020). Modeling time-varying uncertainty of multiple-horizon forecast errors, *Review of Economics and Statistics* **102**(1): 17–33.

Clements, M. (2019). *Maceconomic Survey Expectations*, Palgrave Macmillian, Cham, Switzerland.

Clements, M. P. (2014). Forecasting uncertainty - ex ante and ex post: U.S. inflation and output growth, *Journal of Business and Economic Statistics* **32**: 206–16.

Clements, M. P. (2018). Are macroeconomic density forecasts informative?, *International Journal of Forecasting* **34**(2): 181–198.

Clements, M. P. and Taylor, N. (2001). Bootstrapping prediction intervals for autoregressive models, *International Journal of Forecasting* **17**: 247–267.

D'Amico, S. and Orphannides, A. (2008). Uncertainty and Disagreement in Economic Forecasting, (56/2008).

Engelberg, J., Manski, C. F. and Williams, J. (2008). Comparing the point predictions and subjective probability distributions of professional forecasters, *Journal of Business and Economic Statistics* **27**: 30–41.

Jo, S. and Sekkel, R. (2019). Macroeconomic uncertainty through the lens of professional forecasters, *Journal of Business & Economic Statistics* **37**(3): 436–446.

Knüppel, M. and Krüger, F. (2019). Forecast uncertainty, disagreement, and the linear pool, *Discussion Papers 28/2019*, Deutsche Bundesbank.

Knüppel, M. and Schultefrankenfeld, G. (2019). Assessing the uncertainty in central bank inflation outlooks, *International Journal of Forecasting* **35**(4): 1748–1769.

Krüger, F., Lerch, S., Thorarinsdottir, T. and Gneiting, T. (2020). Predictive inference based on Markov Chain Monte Carlo output, *arXIv* .

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F. and Gneiting, T. (2017). Forecaster's dilemma: extreme events and forecast evaluation, *Statistical Science* **32**: 106–27.

Robertson, J. C., Tallman, E. W. and Whiteman, C. H. (2005). Forecasting using relative entropy, *Journal of Money, Credit and Banking* **37**(383-401).

Smets, F. and Wouters, R. (2007). Shocks and frictions in US business cycles., *American Economic Review* **97**: 586–606.

Turner, D. S. (1990). The role of judgement in macroeconomic forecasting, *Journal of Forecasting* **9**(4): 315–345.

Wallis, K. F. (2004). An assessement of Bank of England and National Institute inflation forecast uncertainties, *National Institute Economic Review* **189**: 64–71.

Wallis, K. F. (2011). Combining forecasts - forty years later, *Applied Financial Economics* **21**(1-2): 33–41.

Wallis, K. F. (2014). The two-piece normal, binormal, or double gaussian distribution: Its origin and rediscoveries, *Statistical Science* **29**(1): 106–112.

Table A1: Evaluating the effect of judgment on the AR densities: "Tilting Approach 1"

Forecast Horizon	Statistical Model	Judgment: moments from professional forecasters							
		Mean				Variance		Skew	
	AR(2)	MPC	NIESR	CE	SEF	MPC	SEF	MPC	SEF
Panel A: Forecasting UK Output Growth (Forecast Origin: 2001Q1-2016Q1)									
	CRPS	Ratios between the CRPS of tilted and original density							
1	0.29	0.99	0.96	0.86		1.63		0.994	
4	0.97	0.85	0.86	0.79	0.86	0.97	1.00	0.996	1.014
8	1.07	1.06	0.99		0.93	0.96	0.99	1.000	1.031
	tw-CRPS	Ratios between the tw-CRPS of tilted and original density							
1	0.07	0.74	1.03	0.90		2.57		1.024	
4	0.19	1.39	1.20	1.22	1.03	**0.54**	**0.51**	0.985	1.051
8	0.17	1.23	1.07		0.93	**0.40**	**0.15**	1.000	1.103
Panel B: Forecasting UK CPI Inflation (Forecast Origin: 2004Q1-2016Q1)									
	CRPS	Ratios between the CRPS of tilted and original density							
1	0.19	**0.61**	1.15	**0.65**		1.04		1.040	
4	0.69	0.94	1.06	0.88	1.03	1.02	1.10	1.017	1.019
8	0.75	1.18	1.09		1.07	**0.94**	1.09	1.012	1.013
	tw-CRPS	Ratios between the tw-CRPS of tilted and original density							
1	0.06	**0.51**	1.19	**0.50**		1.06		1.015	
4	0.20	1.06	1.07	0.65	0.67	**0.83**	0.86	1.028	1.064
8	0.22	1.67	1.28		1.14	**0.53**	**0.26**	1.017	1.032
Panel C: Forecasting Output Growth and Inflation (across forecast horizons and loss functions)									
Prop. Improved		63%	38%	68%	50%	61%	63%	34%	0%
Prop. Sign Improved		11%	9%	18%	0%	34%	38%	3%	0%
Prop. Sign. Worsened		8%	0%	0%	0%	6%	13%	6%	50%

Notes: Tilting Approach 1 imposes one judgment-based predictive moment at a time, keeping the three other moments (up to the 4th moment) at the statistical model values. Values in bold: statistically significant improvement in forecast accuracy due to judgment. Underlined: statistically significant worsening due to judgment. These are based on a two-sided 10% level test of the null of equal forecast accuracy with small sample correction, see eq. (14). The *tw*-CRPS thresholds are $r_1 = 0$ and $r_2 = 4\%$, and these are for year-on-year growth rates. Results for SEF for forecast origins from 2006Q2 onwards only. Panel C: *Prop. Improved* denotes the percentage of occasions (across variables (output growth and inflation, forecast horizons (h=1 to 8) and scores (CRPS and tw-CRPS)) that judgment improves the accuracy of the statistical model's density forecast. *Prop. Sign. Improved* denotes the percentage of occasions that this improvement is statistically significant at the 90% significance level (using individual *t*-tests). *Proportion Sign. Worsened* denotes the percentage of occasions that judgment leads to a statistically significant worsening of forecast performance.

Table A2: Evaluating the effect of judgment on the BVAR densities: "Tilting Approach 1"

Forecast Horizon	Statistical Model	Judgment: moments from professional forecasters							
		Mean				Variance		Skew	
	BVAR	MPC	NIESR	CE	SEF	MPC	SEF	MPC	SEF
Panel A: Forecasting UK Output Growth (Forecast Origin: 2001Q1-2016Q1)									
	CRPS	Ratios between the CRPS of tilted and original density							
1	0.31	0.93	0.90	**0.81**		1.60		1.001	
4	1.02	0.79	**0.80**	**0.73**	**0.82**	0.99	1.00	1.003	1.118
8	1.13	1.00	0.93		0.88	0.97	1.00	1.013	1.184
	tw-CRPS	Ratios between the tw-CRPS of tilted and original density							
1	0.07	0.65	0.90	0.80		2.21		0.998	
4	0.17	1.37	1.15	1.21	1.05	**0.63**	**0.51**	0.995	1.060
8	0.16	0.99	**0.83**		**0.79**	**0.63**	**0.27**	1.033	1.042
Panel B: Forecasting UK CPI Inflation (Forecast Origin: 2004Q1-2016Q1)									
	CRPS	Ratios between the CRPS of tilted and original density							
1	0.20	**0.63**	1.13	**0.67**		1.04		1.002	
4	0.71	0.85	1.00	0.86	0.98	1.00	1.09	1.004	1.000
8	0.79	1.04	1.00		0.98	**0.96**	1.11	1.002	1.004
	tw-CRPS	Ratios between the tw-CRPS of tilted and original density							
1	0.05	0.57	1.18	**0.43**		1.07		1.002	
4	0.22	0.79	0.91	**0.69**	0.59	**0.85**	0.86	1.007	1.014
8	0.24	1.17	1.06		0.92	**0.57**	**0.31**	1.007	1.016
Panel C: Forecasting Output Growth and Inflation (across forecast horizons and loss functions)									
Prop. Improved		66%	63%	78%	88%	69%	50%	14%	0%
Prop. Sign Improved		13%	13%	34%	25%	38%	38%	0%	0%
Prop. Sign. Worsened		0%	0%	0%	0%	3%	13%	22%	38%

Notes: See Notes to Table A1.

Table A3: Evaluating the effect of judgment on the WBS combination densities using the logarithmic score: "Tilting Approach 1"

Forecast Horizon	Statistical Model	Judgment: moments from professional forecasters							
		Mean				Variance		Skew	
	Comb.	MPC	NIESR	CE	SEF	MPC	SEF	MPC	SEF
Panel A: Forecasting UK Output Growth (Forecast Origin: 2001Q1-2016Q1)									
	Log Score	Differences between the Log Score of tilted and original density							
1	-0.75	0.03	0.05	**0.23**		-2.27		-0.01	
4	-3.06	0.10	0.13	0.36	0.03	0.01	-0.37	0.00	-14.0
8	-3.36	-0.15	0.04		**0.18**	-0.08	0.07	-0.23	-17.7
Panel B: Forecasting UK CPI Inflation (Forecast Origin: 2004Q1-2016Q1)									
	Log Score	Differences between the Log Score of tilted and original density							
1	-0.13	-0.31	-2.04	-0.35		-3.14		-0.03	
4	-1.89	-0.18	-0.53	0.33	-2.39	-0.37	-1.32	0.00	-0.05
8	-1.99	-0.22	-0.46		0.18	0.02	-1.73	0.01	-0.03
Panel C: Forecasting Output Growth and Inflation (across forecast horizons and loss functions)									
Prop. Improved		50%	50%	86%	75%	44%	25%	44%	0%
Prop. Sign. Improved		6%	13%	19%	25%	0%	0%	6%	0%
Prop. Sign. Worsened		0%	13%	0%	25%	25%	25%	0%	25%

Notes: See notes to Table A1. The log score statistics are computed having first fitted a kernel to the predictive draws. For the longer horizon forecasts of output growth, during the 2008/2009 recession realizations fell in a region of the forecast density with probabilities near zero, leading to log score values near minus infinity. To minimize the effects of these outliers, in this table we adopt the *ad hoc* strategy of setting all log scores less than -25 to -25. For the differences in the log score, positive values indicate improvements due to judgment.

Table A4: Evaluating the effect of judgment on the WBS combination densities: "Tilting Approach 1" but not constraining the kurtosis (4th moment)

Forecast Horizon	Statistical Model	Judgment: moments from professional forecasters							
		Mean				Variance		Skew	
	Comb.	MPC	NIESR	CE	SEF	MPC	SEF	MPC	SEF
Panel A: Forecasting UK Output Growth (Forecast Origin: 2001Q1-2016Q1)									
	CRPS	Ratios between the CRPS of tilted and original density							
1	0.28	1.03	0.99	0.88		1.26		0.997	
4	0.93	0.85	0.87	0.79	0.88	1.00	1.00	1.001	1.009
8	1.09	1.03	**0.95**		**0.88**	0.99	1.01	1.000	1.011
	tw-CRPS	Ratios between the tw-CRPS of tilted and original density							
1	0.07	0.67	0.99	0.85		1.49		1.000	
4	0.16	1.28	0.97	1.11	0.77	0.88	**0.82**	0.999	1.026
8	0.09	1.41	1.04		0.78	0.91	**0.30**	1.068	1.054
Panel B: Forecasting UK CPI Inflation (Forecast Origin: 2004Q1-2016Q1)									
	CRPS	Ratios between the CRPS of tilted and original density							
1	0.14	**0.84**	1.44	0.83		1.27		1.002	
4	0.68	0.88	1.08	0.87	1.12	0.99	1.04	1.003	1.003
8	0.86	0.98	0.94		0.91	0.97	1.09	1.005	1.008
	tw-CRPS	Ratios between the tw-CRPS of tilted and original density							
1	0.04	0.80	1.06	0.46		1.28		0.996	
4	0.17	0.76	1.08	0.50	0.53	1.06	0.93	0.995	1.008
8	0.20	0.97	**0.82**		0.61	1.06	**0.68**	0.961	0.946
Panel C: Forecasting Output Growth and Inflation (across forecast horizons and loss functions)									
Prop. Improved		66%	72%	86%	88%	38%	50%	44%	13%
Prop. Sign. Improved		3%	9%	13%	13%	3%	38%	3%	0%
Prop. Sign. Worsened		9%	3%	0%	0%	19%	25%	0%	13%

Notes: See notes to Table A1.

Table A5: Evaluating the effect of judgment on the BVAR densities: "Tilting Approach 2"

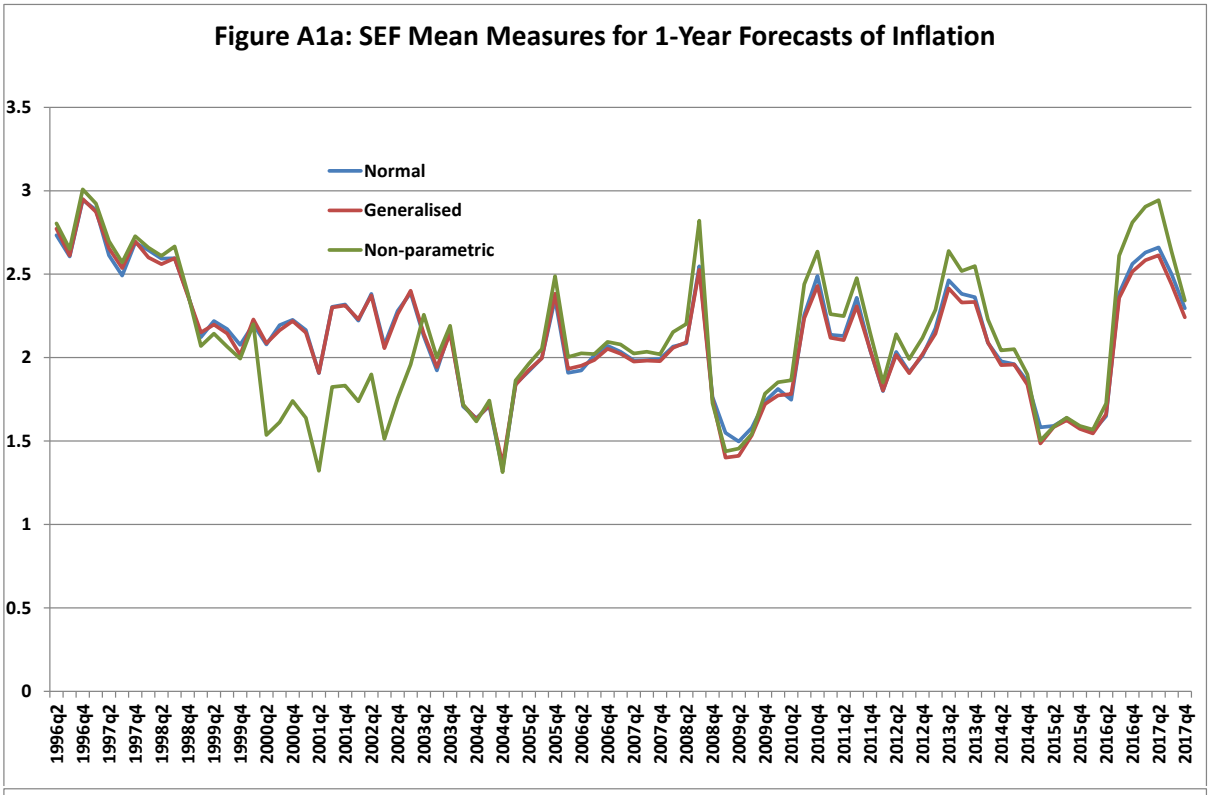
Forecast Horizon	Statistical Model	Judgment: moments from professional forecasters									
		Mean		Variance		Skew		Mean + Var		Mean + Var + Skew	
	BVAR	MPC	SEF	MPC	SEF	MPC	SEF	MPC	SEF	MPC	SEF
Panel A: Forecasting UK Output Growth (Forecast Origin: 2001Q1-2016Q1)											
	CRPS	Ratios between the CRPS of tilted and original density									
1	0.31	0.90		1.15		1.00		1.09		1.21	
4	1.02	0.76	**0.82**	1.00	1.00	0.99	0.99	0.75	0.80	0.76	**0.80**
8	1.13	1.01	0.88	0.98	0.99	0.99	**0.96**	1.00	**0.87**	1.01	**0.87**
	tw-CRPS	Ratios between the tw-CRPS of tilted and original density									
1	0.07	0.61		1.19		**0.99**		0.96		1.23	
4	0.17	1.48	1.06	**0.70**	**0.50**	1.00	**0.92**	1.25	0.69	1.26	0.69
8	0.16	1.08	**0.79**	**0.66**	**0.27**	**0.96**	**0.91**	0.74	**0.15**	0.77	**0.16**
Panel B: Forecasting UK CPI Inflation (Forecast Origin: 2004Q1-2016Q1)											
	CRPS	Ratios between the CRPS of tilted and original density									
1	0.20	**0.57**		1.08		1.00		**0.71**		**0.73**	
4	0.71	0.82	0.98	1.03	1.09	**0.99**	**0.99**	0.85	1.07	0.85	1.07
8	0.79	1.03	0.97	1.02	1.01	**0.99**	0.99	1.08	1.08	1.09	1.08
	tw-CRPS	Ratios between the tw-CRPS of tilted and original density									
1	0.05	0.46		1.10		1.01		0.66		0.65	
4	0.22	0.70	0.58	0.92	0.85	0.99	**0.98**	0.57	**0.22**	0.57	**0.22**
8	0.24	1.15	0.91	**0.66**	0.91	0.99	0.99	0.85	**0.21**	0.86	**0.20**
Panel C: Forecasting Output Growth and Inflation (across forecast horizons and loss functions)											
Prop. Improved		59%	88%	44%	63%	66%	100%	78%	75%	75%	75%
Prop. Sign. Improved		16%	25%	28%	25%	31%	63%	9%	50%	9%	16%
Prop. Sign. Worsened		0%	0%	6%	13%	3%	0%	0%	0%	3%	0%

Notes: In contrast to Tilting Approach 1, Tilting Approach 2 does not constrain other moments at their values from the statistical model (here the BVAR). Also see notes to Table A1.

Table A6: Data Description and Data Transformation for WBS Combination

Name	Description	Transformation BVAR	MIDAS
Target Variables			
Output Growth	GDP expenditure-based method, chained-volume measure, seasonally-adjusted, available on monthly vintages (ONS real-time database).	log	Log-dif
Inflation	CPI Index, All-items, seasonally adjusted, Datastream: UKCONPRCF	log	Log-dif
BVAR with Macro Variables			
Consumption	Final Household Consumption, chained-volume measure, seasonally adjusted. ONS real-time database.	log	
Investment	Gross Fixed Capital Formation, chained-volume measure, seasonally adjusted. ONS real-time database	log	
Real Wages	Unit Labour Cost Index – whole economy, seasonally adjusted. UKLCOST.E. Real values computed using the CPI index.	log	
Hours	Actual hours worked per week, seasonally adjusted. UKYBUS..O	log	
Bank rate	End-of-period Bank of England Base Rate. UKPRATE.	level	
Medium-sized Models			
Industrial Prod.	Index of Production – total manufacturing excluding construction, constant prices, seasonally adjusted. UKIPTOT.G	log	Log-dif
Business Confidence	CBI Monthly Enquiry, Industrial Trends, Volume of Expected Output – Balance - UKCBIOPB	level	Level-dif
Employment	Overall employment, all aged 16 and over, seasonally-adjusted, LFS, UKMGRZ..O	log	Log-dif
Unemployment	Unemployment rate, all aged 16 an over, LFS, UKUN%O16Q	level	level
Consumer Confidence	GFK Consumer Confidence Index, UKGFKCCNR	level	Level-dif
House Prices	LSL/Acadametrics Average House Price in pounds, seasonally-adjusted, UKFTHPI.B	log	Log-dif
Stock Prices	FTSE, all-share index, end of period, UKSHRPRCF	log	Log-dif
Exchange Rates	Real Effective Exchange Rates – CPI based, UKOCC011	log	Log-dif
House Prices _2	Home Sales, output price index. UKPROPRCF	log	log-dif
Short rate	3-month Treasury bills (OECD; UKOIR077R)	level	level
Yield Spread	10-year bond yields (OECD; UKOIR080R) – short rate	level	level
Retail Prices	Retail Price index, all items excluding mortgage interest, not seasonally-adjusted. UKRPAXMIF	log	Log-dif
Oil Prices	Crude Oil prices index (IMF), not seasonally-adjusted. WDI76AADF	log	Log-dig

Figure A1: Alternative Methods to Estimate the Mean and the Variance of the Bank of England Survey of External Forecasters histograms for one-year-ahead Inflation Forecasts.



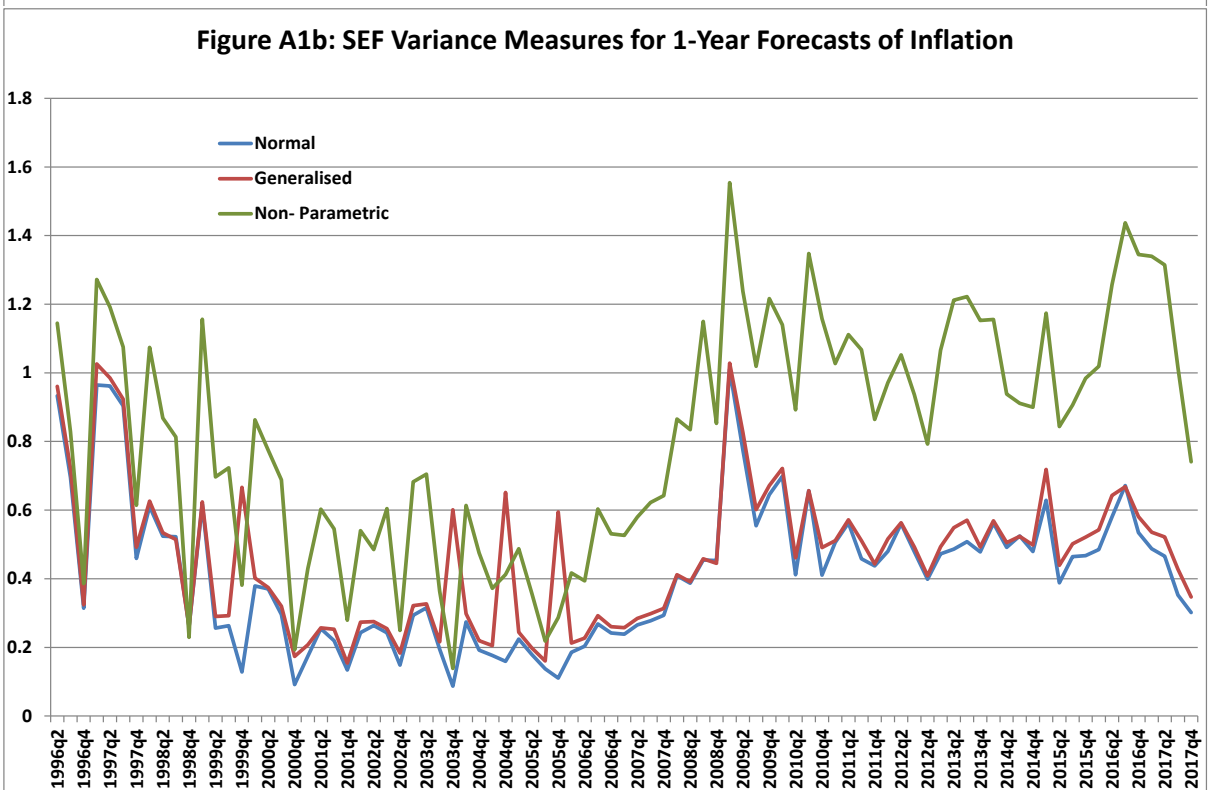
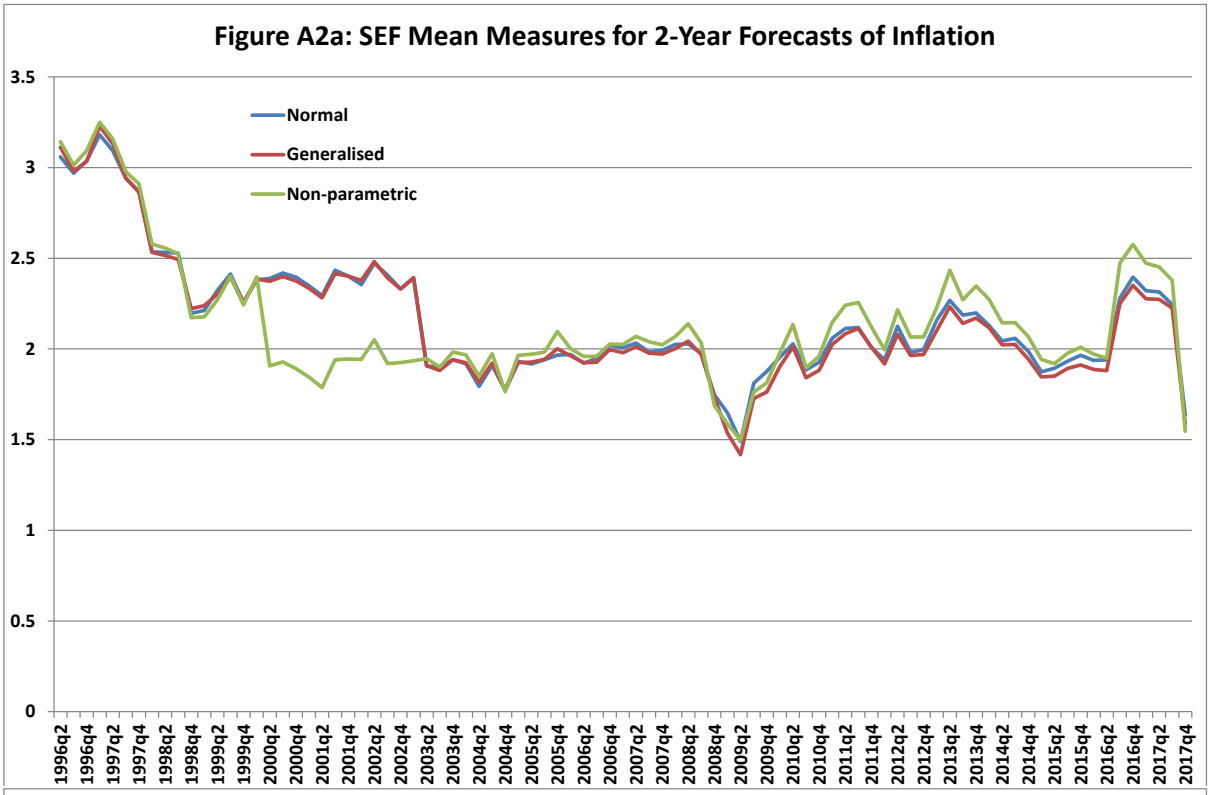


Figure A2: Alternative Methods to Estimate the Mean and the Variance of the Bank of England Survey of External Forecasters histograms for two-year-ahead Inflation Forecasts.



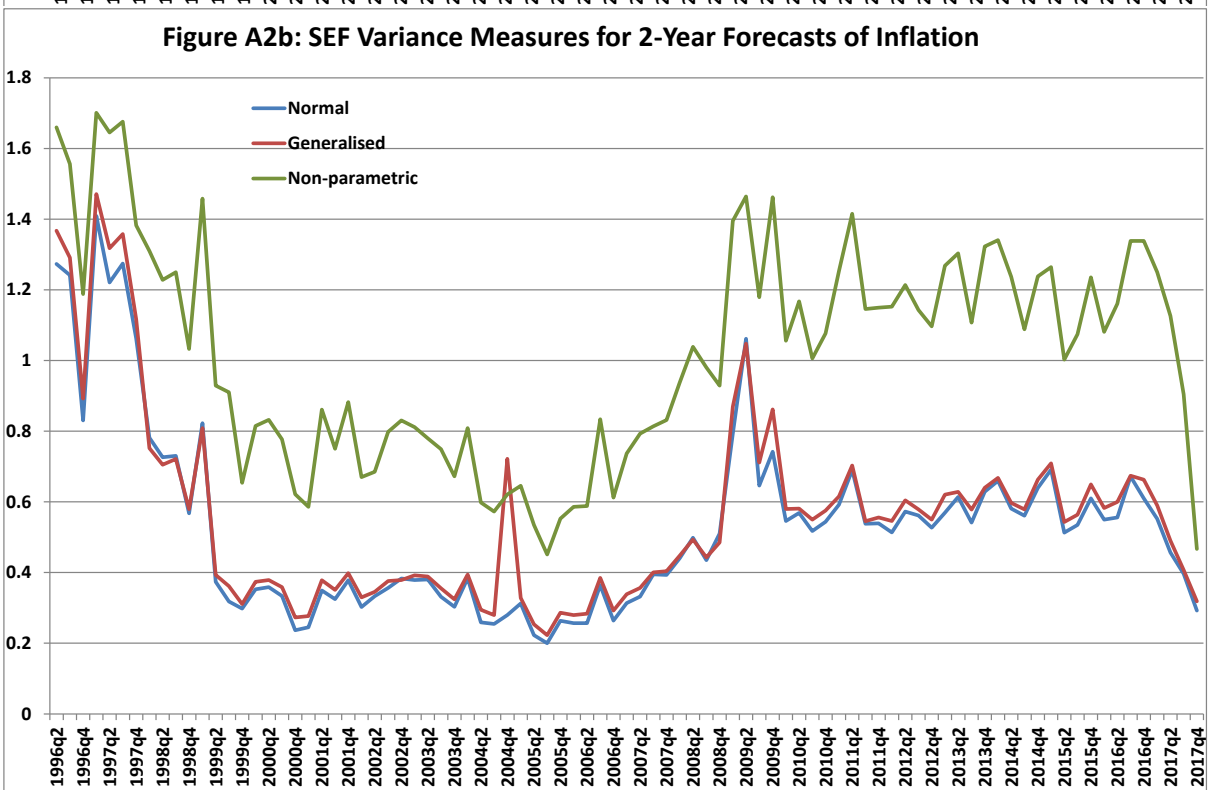
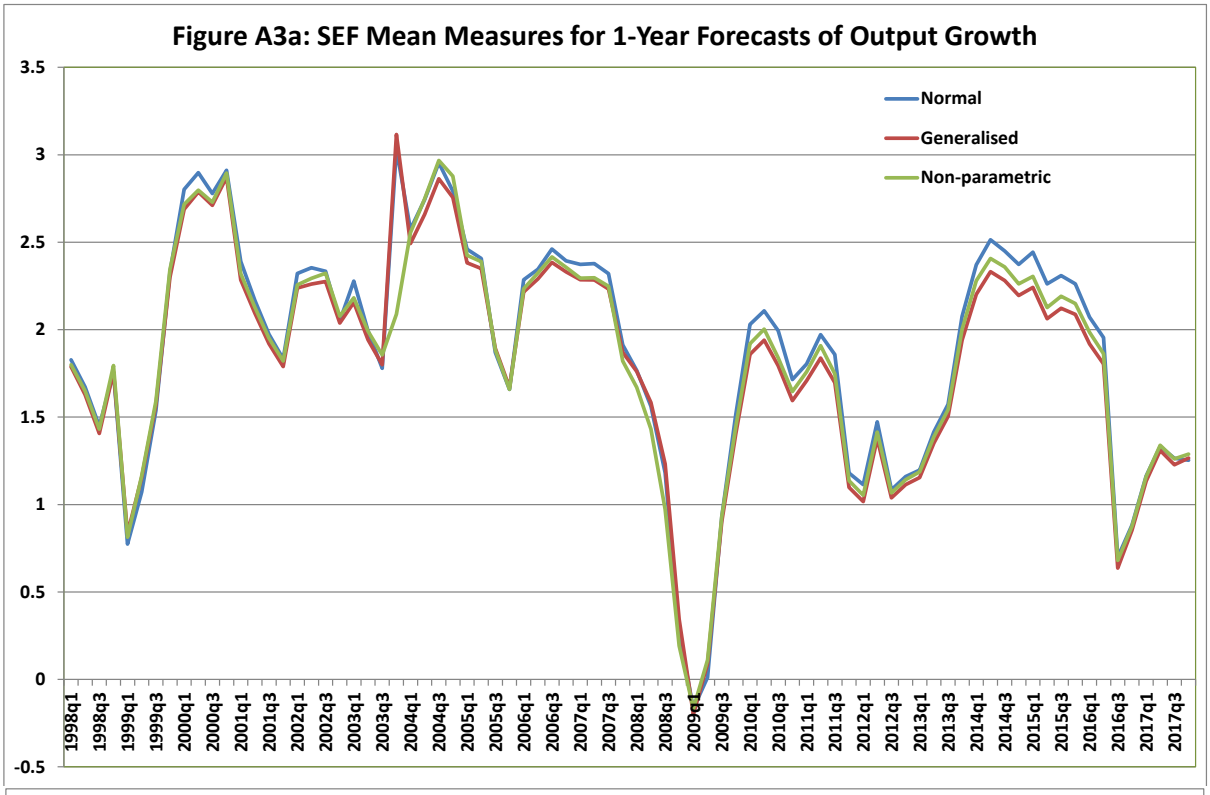


Figure A3: Alternative Methods to Estimate the Mean and the Variance of the Bank of England Survey of External Forecasters histograms for one-year-ahead Output Growth Forecasts



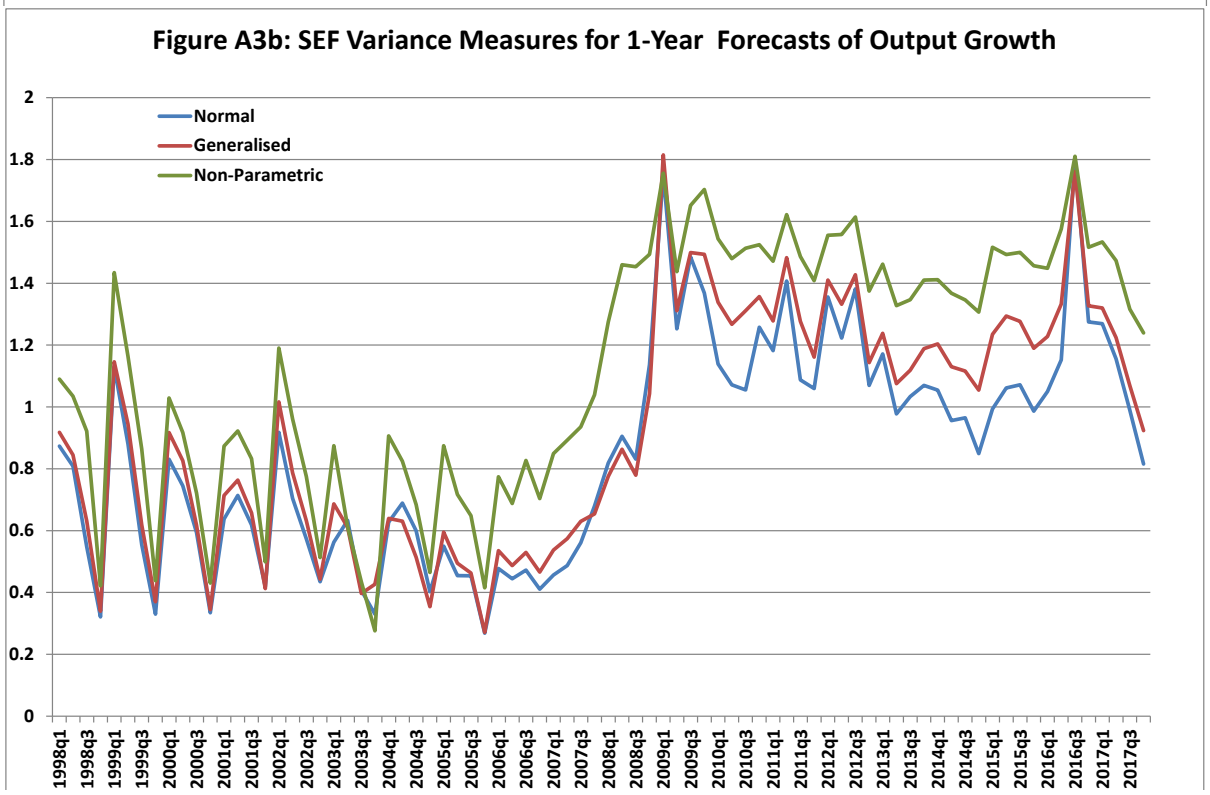
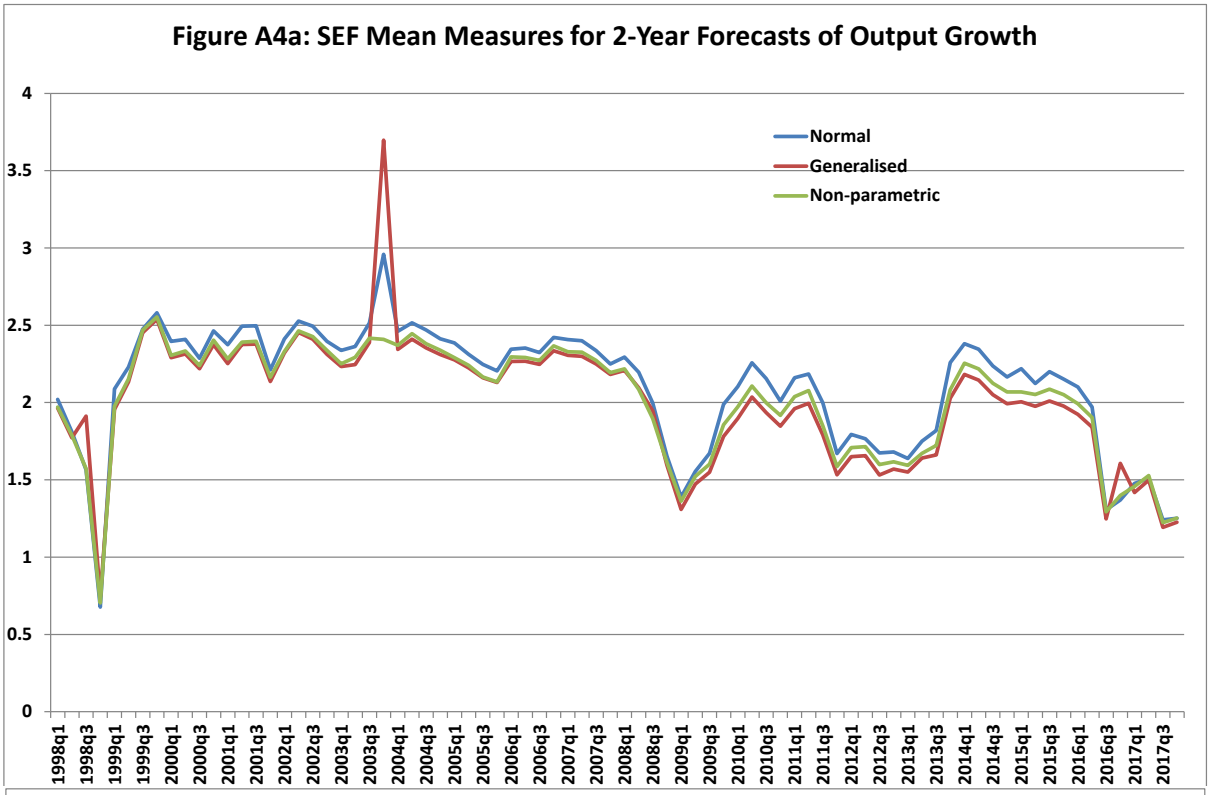


Figure A4: Alternative Methods to Estimate the Mean and the Variance of the Bank of England Survey of External Forecasters histograms for two-year-ahead Output Growth Forecasts



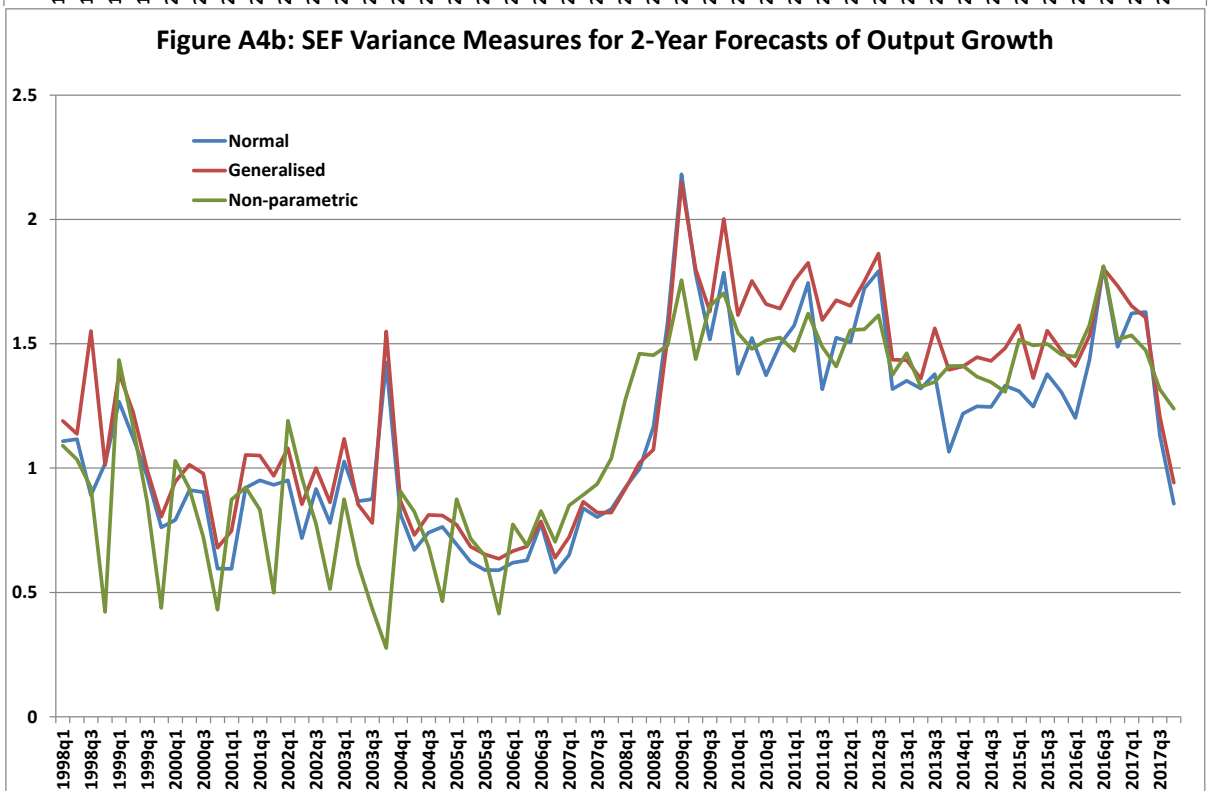


Figure A5: Predicted uncertainty (standard deviation) for MPC: ex post (or implicit, "imp") vs ex post (or explicit "exp") variance forecasts for output growth

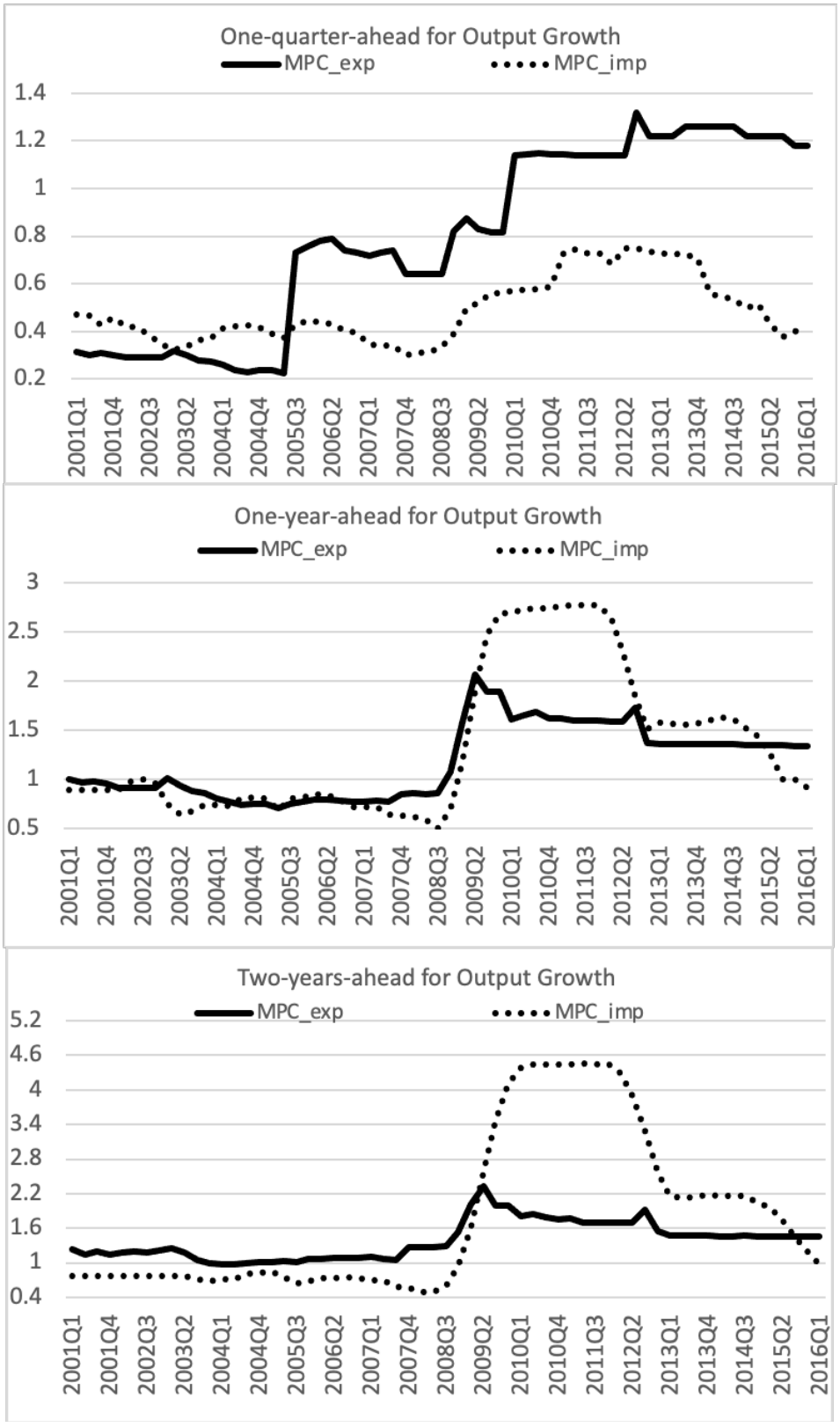


Figure A6: AR, BVAR and WBS combination forecasted skew for the two-year-ahead forecasts

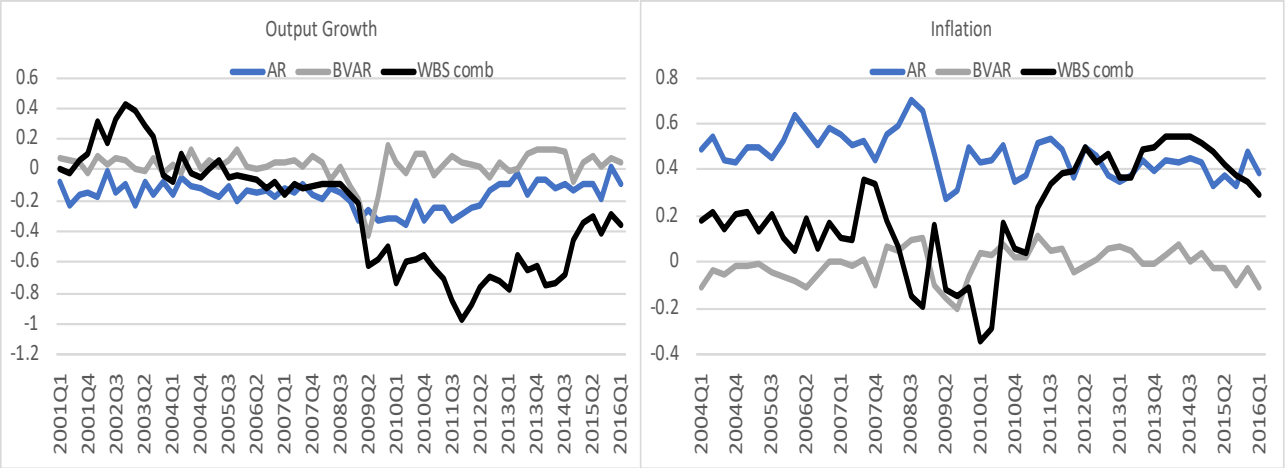
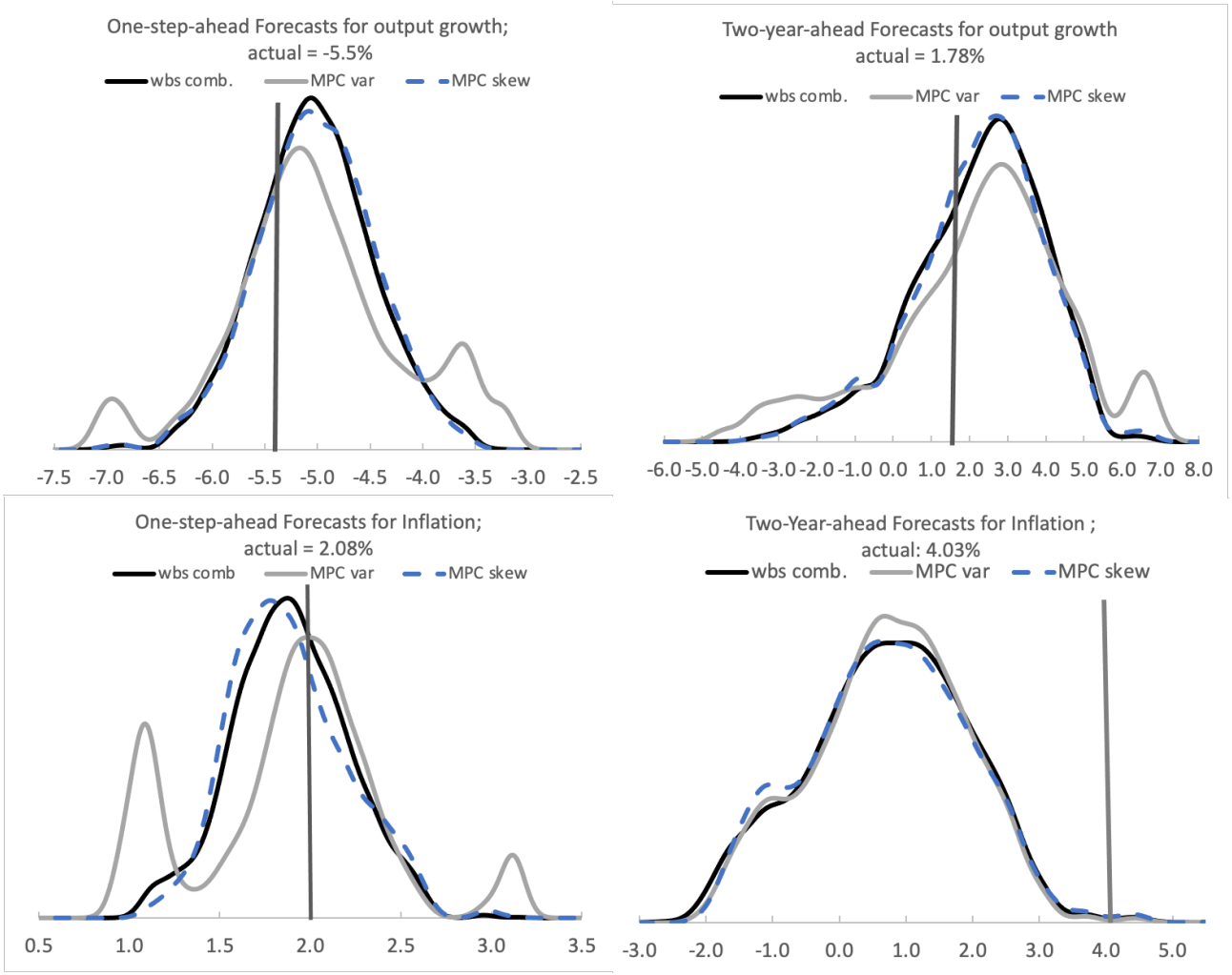


Figure A7: Illustrating the effects of tilting the WBS combination density forecasts towards the MPC's variance and skew forecasts made in 2009Q2 using "Tilting Approach 1"



Notes: Each panel plots the original (WBS combination) and two tilted density forecasts, the first tilting towards the MPC's variance forecast and the second tilting towards the MPC's skew forecast. The densities are estimated by fitting a kernel (normal with a fixed bandwidth) to the draws from the original and tilted densities. The solid vertical line is the realization of the variable.

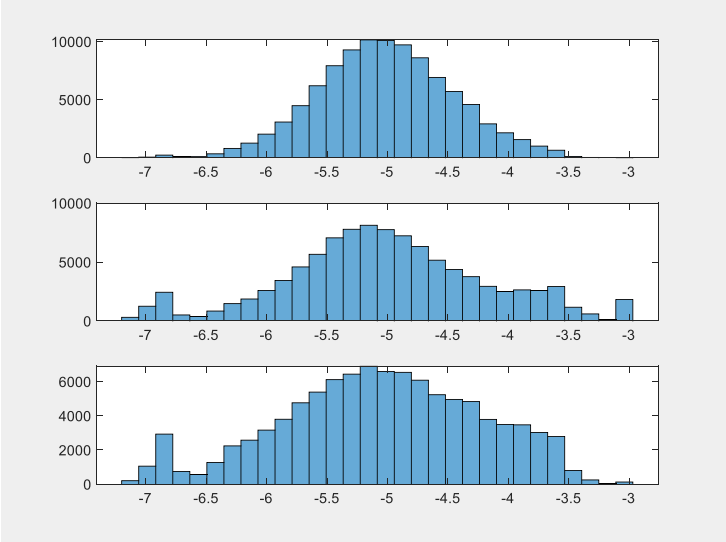
Top left panel: the first four moments of the WBS combination forecast are -5.03, 0.31, 0.05 and 2.97. The MPC's variance forecast is 0.69 and their skew forecast is -0.13.

Top right panel: the first four moments of the WBS combination forecast are 2.23, 2.86, -0.61 and 3.4. The MPC's variance forecast is 5.4 and their skew forecast is -0.54.

Bottom left panel: the first four moments of the WBS combination forecast are 1.89, 0.10, 0.09 and 3.00. The MPC's variance forecast is 0.26 and their skew forecast is 0.45.

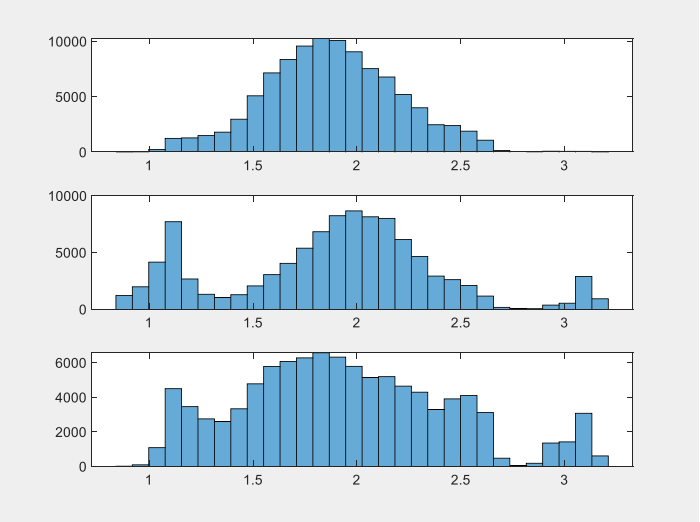
Bottom right panel: the first four moments of the WBS combination forecast are 0.73, 1.47, -0.10 and 2.58. The MPC's variance forecast is 1.33 and their skew forecast is 0.07.

Figure A8: Illustrating the effects of tilting the WBS combination density forecasts towards the MPC's variance forecasts made in 2009Q2 using "Tilting Approach 1": one-quarter-ahead output growth



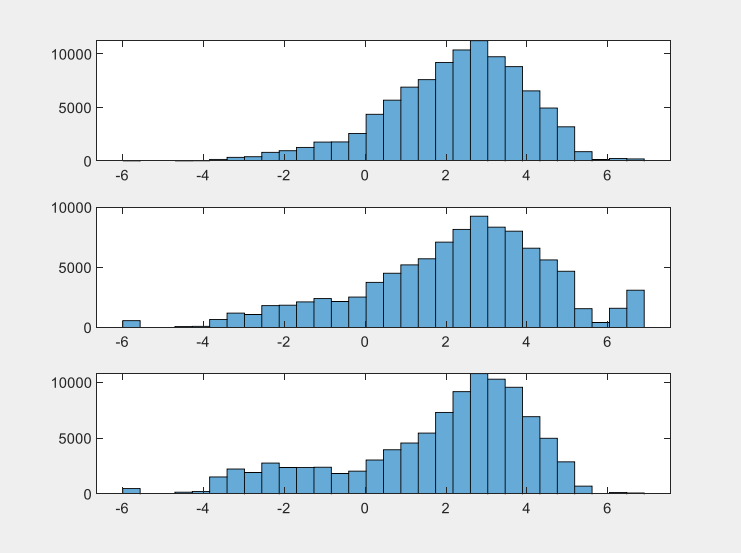
Notes: Histograms for original and tilted samples (D=100,000).
Top panel: WBS combination forecast.
Middle panel: is the tilted density, tilting towards the MPC variance forecast of 0.69 keeping the other first three moments unchanged. The first four standardized moments of the original forecast are -5.03, 0.31, 0.05 and 2.97.
Bottom panel: is the tilted density, tilting towards the MPC variance forecast of 0.69 but not constraining the other first three moments to remain unchanged. The first four standardized moments of the tilted forecast are -5.03, 0.69, -0.22 and 2.58.

Figure A9: Illustrating the effects of tilting the WBS combination density forecasts towards the MPC's variance forecasts made in 2009Q2 using "Tilting Approach 1": one-quarter-ahead Inflation



Notes: Histograms for original and tilted samples (D=100,000).
Top panel: WBS combination forecast.
Middle panel: is the tilted density, tilting towards the MPC variance forecast of 0.26 keeping the other first three moments unchanged. The first four standardized moments of the original forecast are 1.89, 0.10, 0.09 and 3.00.
Bottom panel: is the tilted density, tilting towards the MPC variance forecast of 0.26 but not constraining the other first three moments to remain unchanged. The first four standardized moments of the tilted forecast are 1.95, 0.26, 0.37 and 2.69.

Figure A10: Illustrating the effects of tilting the WBS combination density forecasts towards the MPC's variance forecasts made in 2009Q2 using "Tilting Method 1": two-year-ahead output growth



Notes: Histograms for original and tilted samples (D=100,000).
Top panel: WBS combination forecast.
Middle panel: is the tilted density, tilting towards the MPC variance forecast of 5.4 keeping the other first three moments unchanged. The first four standardized moments of the original forecast are 2.23, 2.86, -0.61 and 3.40.
Bottom panel: is the tilted density, tilting towards the MPC variance forecast of 5.4 but not constraining the other first three moments to remain unchanged. The first four standardized moments of the tilted forecast are 1.81, 5.35, -0.91 and 3.15.