Working Paper

Macro modelling with many models

by Ida Wolden Bache, James Mitchell, Francesco Ravazzolo and Shaun P. Vahey

%NB% NORGES BANK

Macro Modelling with Many Models*

Ida Wolden Bache[†]   James Mitchell[‡]   Francesco Ravazzolo[§]

(Norges Bank)   (NIESR)   (Norges Bank)

Shaun P. Vahey[¶]

(Melbourne Business School)

August 17, 2009

Abstract

We argue that the next generation of macro modellers at Inflation Targeting central banks should adapt a methodology from the weather forecasting literature known as 'ensemble modelling'. In this approach, uncertainty about model specifications (e.g., initial conditions, parameters, and boundary conditions) is explicitly accounted for by constructing ensemble predictive densities from a large number of component models. The components allow the modeller to explore a wide range of uncertainties; and the resulting ensemble 'integrates out' these uncertainties using time-varying weights on the components. We provide two examples of this modelling strategy: (i) forecasting inflation with a disaggregate ensemble; and (ii) forecasting inflation with an ensemble DSGE.

Keywords: ensemble modelling, forecasting, DSGE models, density combination

JEL codes: C11; C32; C53; E37; E52

---

1 Introduction

We argue that macro models in Inflation Targeting countries are too narrowly focused to forecast probabilities well. Despite the explicit consideration of model uncertainty afforded by Bayesian estimation techniques, the models prominent in central banks devote insufficient attention to 'uncertain instabilities'. That is, too much attention has been paid to refining a single preferred but inevitably misspecified model. A product of this oversight is that the 2007 vintage workhorse monetary policy models had little (or nothing) to say about the probability of 'tail' events which now dominate the debate over the causes of, and remedies for, the recent Global Financial Crisis.

In our view, the next generation of macro modellers should address this deficiency whilst preserving the architecture of dynamic non-linear modelling. We propose a methodology adapted from the weather forecasting literature known as 'ensemble modelling' . In this approach, uncertainty about model specifications (e.g., initial conditions, parameters, and boundary conditions) are explicitly accounted for by constructing ensemble predictive densities from a large number of component models. The components allow the modeller to explore a wide range of uncertainties; and the resulting ensemble 'integrates out' these uncertainties using time-varying post-data weights on the components.

We provide two economic examples of the ensemble methodology. In the first, we consider a policymaker (recursively) selecting a linear combination of disaggregate predictives to produce an ensemble forecast density for inflation based on disaggregate measures of inflation. Each component of the ensemble comprises a univariate autoregessive model using a single disaggregate series. In our second application, we utilize an ensemble of DSGE models, where the components are differentiated by candidate break dates. In both examples, the ensembles outperform autoregressive benchmarks in terms of density forecast performance for Norwegian inflation.

The remainder of this paper is structured as follows. In the next section, we set out the case for explicit consideration of ensemble methods to deal with uncertain instabilities. In section 3, we provide an example of ensemble modelling based on combining the information in disaggregate inflation indices; and a second example combines DSGE models. In the final section, we conclude by considering the scope for further developments in macro modelling under Inflation Targeting.

2 Current macro and ensemble modelling

We begin this section by motivating our approach to macro modelling. Then we outline the relationship between ensemble modelling and the 'uncertain instabilities' literature in macroeconomics. We complete the section with a discussion of the characteristics of the ensemble approach.

2.1 What just happened?

The recent Global Financial Crisis has provoked considerable debate about the nature of existing macro models. In particular, it has been argued that the models at the heart of policymaking in Inflation Targeting central banks abstract from key aspects of financial plumbing that were the source of the crisis. And that without these features, modellers and policymakers stood little chance of spotting the slump in activity that stemmed from it. For example, Willem Buiter ('Maverecon', Financial Times blog, March 3) has argued that modern macro "excludes everything I am interested in" , echoing the thoughts of Charles Goodhart on the dominance of macro models without finance (e.g., see Goodhart, 2007).

In response, central bankers are, no doubt, busy bolting bits of financial apparatus onto workhorse DSGE models, which abstract currently from many key features of the economy, not just the financial plumbing implicated in the current crisis. The narrow focus on inflation above other considerations represents, after all, the defining feature of an Inflation Targeting regime. Leaving aside the issue of whether macro models should be developed by bolting on new sectors after each unique crisis, whatever bits are added to today's model, the next generation of central bank workhorse models will remain highly abstract.

Of course, the early RBC literature—which kick-started the computational developments prevalent in modern workhorse macro models—gave explicit consideration to the extreme degree of abstraction. Given this starting point, it seems surprising to us that Inflation Targeting central banks typically focus on a single specification, estimated directly by the Bayesian equivalent of Maximum Likelihood.[1] Why should simply allowing for informative priors (on parameters) yield a specification capable of producing accurate forecasts for the events of interest to policymakers?

---

[1] Karagedikli et al (2009) provide a recent review of DSGE modeling.

2.2   Uncertain instabilities and ensemble modelling

A recent strand of the macro-econometrics literature has focused on 'model uncertainty' more widely, taking as its foundation that the models considered are profoundly false. Durlauf and Vahey (2009) provide a summary in a special issue of the *Journal of Applied Econometrics* which focuses on the approach. On the forecasting side, this framework is sometimes referred to as one of 'uncertain instabilities'. The name implies that the estimated parameters of a single model will exhibit instabilities and that these can be difficult to identify in the real-time forecasting exercises confronting central banks. The dominant strategy in the forecasting applications is to combine the evidence from many models. For example, Clark and McCracken (2009) examine the scope for taking linear combinations of point forecasts in real time, motivated by the desire to circumvent the uncertain instabilities in any particular specification.[2] In a series of papers, Stock and Watson (2001, 2004) have documented the robust performance of point forecast combinations using various types of models for numerous economic and financial variables.

In so far as the 'uncertain instabilities' literature combines the evidence from many specifications, the prevailing approach has a Bayesian interpretation. The difficulty with applying conventional frequentist econometrics here is obvious: selecting a single model has little appeal if the usual model selection approach yields a specification that suffers from instability. This might happen either if the 'true' model is not within the model space considered by the modeller, or if the model selection process performs poorly on short runs of macroeconomic data.

Geweke (p95, 2009) argues that standard Bayesian methods are ill-suited to the tasks of inference and prediction in the case where the 'true' model is absent from the model space—sometimes referred to as an 'incomplete model space'. A number of econometricians, including Sims (2007) and Del Negro *et al.* (2007) have noted that ratios of marginal likelihoods overstate the difference between candidate models in the absence of the 'true' model from the model space. Using several examples, Geweke (ch 5, 2009) demonstrates the scope for pooling forecast densities to produce superior predictions, even if the set of components to be combined excludes the 'true' model. Hall and Mitchell (2007) draw attention to this property in their analysis of forecast density performance and combination at two institutions, namely, the Bank of England and the National Institute of Economic and Social Research.

Outside the econometrics literature, the benefits of density combination have been

---

[2]Jore, Mitchell and Vahey (2009) examine linear combinations of densities using the same US data.

recognized for some time, as Garratt, Vahey and Mitchell (2009) observed. Over the last 15 years, meteorologists and statisticians have focused a great deal of attention on analyzing statistical ensembles. The roots of the approach can be traced back to Gibbs' contribution to thermodynamics. Loosely, the idea behind the ensemble approach is to consider a large number of component models, each of which is a replicant of the 'preferred' specification.[3] Since each component could be viewed as an approximation of the current state of the 'true' but unknown specification, the components are considered together. The ensemble of components approximates the truth.

In the meteorological forecasting literature, the ensemble methodology is a response to the 'uncertain instabilities' problem. Density forecasts are generated from a common theoretical framework with slightly different initial conditions (measurements, auxiliary assumptions). The framework from which the component specifications are derived might allow for data, parameter, and/or model uncertainty.[4] With practical forecasting issues in mind, two central questions are: (i) What components should be included in the model space?; and (ii) How should the ensembles be simulated? The researcher designs the ensemble component model space in order to explore the likely source of 'uncertain instabilities'. In meteorology and climate prediction, the analysis typically uses Monte Carlo simulation techniques.

Ensemble predictive methods are commonly used by the majority of weather prediction institutions worldwide. One example is the "Ensemble Prediction System" developed by the European Centre for Medium-Range Weather Forecasts.[5] Leutbecher and Palmer (1997) provide a primer on ensemble forecasting in meteorology. MacKenzie (2003) considers the impetus to ensemble developments in meteorology provided by failing to assess the probability of severe storms (tail events).

The experience of the Global Financial Crisis indicates the limitations of designing monetary policy in the absence of precise information about probabilities. The ensemble methodology offers the scope to generate accurate density forecasts from large numbers of theoretically-coherent models. Even with explicit consideration of financial plumbing, it is hard to envisage that a single next-generation DSGE model will offer accurate (and robust) probabilistic forecasts.

---

[3]The term 'replicant' is taken from the movie 'Blade Runner' (1982). The equivalent term in micro is 'differentiated but otherwise identical'.

[4]In weather forecasting applications, the sensitivity to initial conditions stems from the chaotic processes considered.

[5]For an early description of weather ensemble forecasting see Molteni *et al.* (1996).

2.3 Characteristics of ensemble modelling

We conclude this section by highlighting some common characteristics of an ensemble modelling strategy for macro modelling.

1. Generation of forecasting densities, rather than point forecasts

2. Predictive density construction from a large number of component macro-econometric models

3. Forecast density evaluation and combination based on out of sample performance, rather than in-sample analysis

4. Component model weights vary through evaluation—ensemble densities have time varying weights

Papers in the economics literature that satisfy these criteria include (among others): Jore, Mitchell and Vahey (2009), Kasha and Ravazzolo (2009), Gerard and Nimark (2008) and Garratt, Mitchell and Vahey (2009). Smith *et al.* (2009) consider the performance of the Norges Bank nowcasting system which also adopts the ensemble methodology. In these cases, the out of sample densities from many macro-econometric component models are directly combined into the ensemble using an 'opinion pool'.[6] These papers differ in the design of the model space and the number of components considered, as well as the applied problem of interest. We shall discuss this opinion pool approach in considerable detail below when we analyze two specific examples. As we shall discuss, variants can produce symmetric or non-symmetric predictive densities.

Another strand of the ensemble economics literature uses informative priors and Markov chain Monte Carlo methods to produce ensembles. Maheu and Gordon (2009) and Geweke (2009) use mixture models to give non-Gaussian predictives; Andersson and Karlsson (2007) produce symmetric Gaussian predictive densities from many vector autoregressions.[7] Geweke (2009) discusses the relationships between density pooling and mixture modelling, and argues that the former presents a more coherent approach for incomplete model spaces. Clearly, both variants can be effective methods for combining densities in forecasting applications. (In a related literature, Patton (2004), Maheu and

---

[6]Wallis (2005) uses opinion pools to average (model free) survey forecasts, rather than those from macro-econometric models. Mitchell and Hall (2005) use opinion pools to combine forecasts from two institutions.

[7]Frequentist approaches to mixture model estimation are also feasible but practitioners have tended to prefer Bayesian simulation methods with scope for informative priors.

McCurdy (2009) and Amisano and Geweke (2009) consider ensembles in various financial applications.)

The mixture innovations approach to state space models developed by (for example) Giordani, Kohn and van Dijk (2007) and Giordani and Kohn (2008) has a number of common features with ensemble modelling. Both strategies aim to combine relatively simple components. In ensemble applications, the components are typically conditionally (i.e., locally) linear Gaussian, but this is not required. The mixture innovation literature deploys the Kalman filter to conditionally linear Gaussian processes.[8] Given the flexibility afforded by combination, the Gaussian components may not impair forecasting performance significantly. Ensemble modelling applications also focus more explicitly on out of sample density forecasting, and given the relatively light computational burden, a broader (and sometimes more eclectic) model space made up of candidate macro-econometric specifications.

3   Examples

In this section, we consider two specific examples of ensemble forecasting for inflation: using an ensemble of disaggregates, and DSGE ensembles. Both applications use Norwegian data and the opinion pool approach to ensemble density construction. We begin by describing the density combination approach used throughout.

3.1   Density combination

To summarize the approach, for each observation in the policymaker's out of sample 'evaluation period', we use forecast density combination to compute the weight on each component model. In each example that follows, the component models will use a common time series structure. In the first example, each component will consider a particular disaggregate inflation measure. In the second example, the DSGE components will be distinguished by assumptions about break date timing. In both cases, the weights are based on the 'fit' of the component predictive densities for measured inflation. Given these weights, we construct ensemble forecast densities for measured inflation.

More formally, consider a policymaker aggregating forecasts supplied by 'experts' each using a unique component forecasting model. Given $i = 1, \ldots, N$ components (where N

---

[8]The ensemble Kalman filter can be used to approximate non-Gaussian processes with an ensemble based on simulated Gaussian measurement errors; see, for example, Mandel (2007).

could be a large number), we define the ensemble by the convex combination also known as a linear opinion pool:

$$p(\pi_{\tau,h}) = \sum_{i=1}^{N} w_{i,\tau,h} \; g(\pi_{\tau,h} \mid I_{i,\tau}), \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}, \tag{1}$$

where $g(\pi_{\tau,h} \mid I_{i,\tau})$ are the h-step ahead forecast densities for inflation from component model i, $i = 1, \ldots, N$, conditional on the information set I_{τ}.

We stress that each component model produces h-step ahead forecasts for inflation— we are weighting the components only by the performance of their inflation forecasts in these examples, although multivariate weights are feasible.[9]

Each component model uses data, dated $\tau - h$ or earlier, to produce an h-step ahead forecast density for τ. The non-negative weights, $w_{i,\tau,h}$, in this finite mixture sum to unity, are positive, and vary by recursion in the evaluation period $\tau = \underline{\tau}, \ldots, \overline{\tau}$.

We emphasize that the ensemble forecast density could be non-Gaussian even if the component models produce Gaussian predictives. The linear opinion pool ensemble (1) accommodates skewness and kurtosis. The flexible structure resulting from linear pooling allows the data to reveal whether, for example, the ensemble should have fat tails, or asymmetries. Kascha and Ravazzolo (2009) compare and contrast logarithmic and linear pooling. Logarithmic opinion pools force the ensemble predictives to be symmetric, but accommodate fat tails; see also, Smith *et al.* (2009).

We construct the ensemble forecast density for measured inflation using equation (1). Implementation of the density combination requires a measure of component density fit to provide the weights. A number of recent applications in the economics literature have used density scoring rules. In the applications that follow, we utilize the Continuous Ranked Probability Score (CRPS), which as (among others) Gneiting and Raftery (2007), Panagiotelis and Smith (2008) and Ravazzolo and Vahey (2009) note, rewards predictive densities from component models with high probabilities near (and at) the outturn.[10]

3.2   A disaggregate ensemble

Monetary policymakers examine regularly disaggregate inflation series for leading evidence of the inflationary process. The introduction of Inflation Targeting led central banks to

---

[9]In the examples that follow, we set $h = 1$ for simplicity.

[10]See Panagiotelis and Smith (2008) for an explanation of how CRPS is calculated for each component density.

focus much greater attention on the behaviour of inflation. One problem in doing so is that headline inflation can be volatile. A tradition common among Inflation Targeters considers the disaggregate inflation (or price) cross-sectional distribution but truncates and averages the distribution to provide a 'core' measure. A second approach excludes (zero-weights) particular disaggregates; the resulting measure is commonly referred to as an 'Ex' core measure. It is often argued that a key test of core inflation measures is whether the candidate core can predict measured inflation at a given horizon; see, for example, Smith (2004).

In this example, we construct ensemble predictives based on the out of sample forecast performance of many component models, where each component model uses a particular disaggregate series. The example follows closely the approach of Ravazzolo and Vahey (2009). Using US data, they label the combined forecast density the 'disaggregate ensemble core' inflation. We demonstrate below that the ensemble predictives provide accurate forecast densities for measured inflation. And the weights on the disaggregate components are non-zero throughout the evaluation. We conclude that the common practice of discarding disaggregate information either by zero-weighting groups or individual disaggregates—analogous to truncation and the 'Ex' approach, respectively—are unwarranted from the perspective of assessing the probability of inflation events of interest. The example also illustrates the mechanics and flexible nature of ensemble modelling.

In our application, we decompose inflation in Norway into 12 disaggregates. These are: food and non-alcoholic beverages; alcoholic beverages and tobacco; clothing and footwear; housing, water, electricity and fuels; furnishings and house equipment; health care; transport; communications; recreation; education; restaurants and hotels; and, miscellaneous goods and services. We emphasize that, in principle, our methodology could be applied to an extremely large number of disaggregates. For all inflation series, we work with quarterly growth rates. Restricting our attention to Great Moderation data, we start our sample with 1984Q1 and end with 2008Q4. The evaluation period for the predictives is 1996Q3 to 2008Q4; the period 1993Q1 to 1996Q2 we use as a 'training period' to initialize the ensemble weights. This application focuses entirely on one-step ahead forecasts.[11]

Recall that we construct the ensemble by combining the predictive densities from

---

[11]Within the core inflation literature, the horizon of interest varies, typically between one and eight quarters ahead. Although longer horizon ensemble forecasts are possible (see, for example, Jore, Mitchell and Vahey (2009)), we prefer to focus on horizons much shorter than the focal range of many Inflation Targeting regimes. Thereby, the results presented in this paper cannot be interpreted as a test of the 'credibility' of the inflation targeting regime. For further discussion of this issue see Brischetto and Richards (2006).

all of the disaggregate component models. Each component model uses a univariate autoregressive specification with four lags for a single disaggregate series. We construct the ensemble predictives for measured inflation by evaluating the disaggregate forecasts for measured inflation. In each recursion, we (recursively) center the component forecasts on measured inflation as described by Ravazzolo and Vahey (2009).[12]

To assess the calibration properties of the core ensemble density we follow Diebold *et al.* (1998) and compute *PITS*, probability integral transforms, and apply the Berkowitz (2001) likelihood ratio test for independence, zero mean and unit variance of the *PITS*. The test statistic is distributed $\chi^2(3)$ under the null hypothesis of no calibration failure, under a maintained hypothesis of normality. We also report the average (over the evaluation period $T = \bar{\tau} - \underline{\tau}$) logarithmic score. The logarithmic score of the i-th density forecast, $\ln g(\pi_{\tau,h} \mid I_{i,\tau})$, is the logarithm of the probability density function $g(. \mid I_{i,\tau})$, evaluated at the outturn $\pi_{\tau,h}$. Hence, the log score evaluates the predictives at the outturn only. We investigate relative predictive accuracy by considering a Kullback-Leibler information criterion (KLIC)-based test, based on the expected difference in two models' log scores; see Bao *et al.* (2007), Mitchell and Hall (2005) and Amisano and Giacomoni (2007). Suppose there are two density forecasts, $g(\pi_{\tau,h} \mid I_{1,\tau})$ and $g(\pi_{\tau,h} \mid I_{2,\tau})$, so that the KLIC differential between them is the expected difference in their log scores: $d_{\tau,h} = \ln g(\pi_{\tau,h} \mid I_{1,\tau}) - \ln g(\pi_{\tau,h} \mid I_{2,\tau})$. The null hypothesis of equal density forecast accuracy is $\mathcal{H}_0 : E(d_{\tau,h}) = 0$. A test can then be constructed since the mean of $d_{\tau,h}$ over the evaluation period, $\bar{d}_{\tau,h}$, under appropriate assumptions, has the limiting distribution: $\sqrt{T}\bar{d}_{\tau,h} \to N(0, \Omega)$, where Ω is a consistent estimator of the asymptotic variance of $d_{\tau,h}$.[13] Mitchell and Wallis (2009) explain the importance of information-based methods in discriminating between competing density forecasts.

We construct an ensemble one-step ahead predictive density for measured inflation, which we refer to as DE12. As a benchmark, we use a linear model to forecast measured inflation without disaggregate information. That is, we use a linear autoregressive model for aggregate measured inflation, with four lags, AR(4).[14] We use this AR model as our benchmark in tests of relative forecast performance.

---

[12]In effect, this step restricts the ensemble forecast densities to be uni-modal but not symmetric.

[13]When evaluating the density forecasts we treat them as primitives, and abstract from the method used to produce them. Amisano and Giacomoni (2007) and Giacomini and White (2006) discuss more generally the limiting distribution of related test statistics.

[14]We use uninformative priors for the AR(4) parameters with an expanding window. The predictive densities follow the t-distribution, with mean and variance equal to OLS estimates; see, for example, Koop (2003) for details.

Table 1: Forecast performance

	LR	LS	LS-test
AR	0.175	-1.057	
DE12	0.215	-0.615	0.001

Note: The column LR is the Likelihood Ratio p-value of the test of zero mean, unit variance and independence of the inverse normal cumulative distribution function transformed *PITS*, with a maintained assumption of normality for transformed *PITS*. LS is the average logarithmic score, averaged over the evaluation period. LS-test is the p-value of the KLIC-based test for equal density forecasting performance of AR and DE12 over the sample 1996Q3 to 2008Q4.

Before turning to the density evaluations for our various ensembles, we summarize the point forecast performance. The root mean squared prediction error (RMSPE) of DE12 and AR(4) is 0.313 and 0.430, respectively. The Clark-West (2006) test for superior predictive accuracy (against the null of equal accuracy) indicates the superior performance of DE12 with a test statistic of 2.61; the critical value for rejection of the null at 95% is 1.65.[15]

We turn now to the ex post (end of evaluation period) evaluation of the forecast densities from DE12 and the AR(4) benchmark. Table 1 has two rows; one for each. The columns report (reading from left to right) the Berkowitz likelihood ratio test (based on the *PITS*), the log scores (averaged over the evaluation period), and the *p*-values for the equal predictive density accuracy test (based on the log scores), respectively. Whereas both models appear well-calibrated on the basis of the Berkowitz likelihood ratio, the final column shows that the AR is rejected in favour of DE12 using the KLIC-based test. DE12 delivers a statistically significant, at the 99% level, improvement in the log score (reported in the second column).

The weights in DE12 display some variation through time. Table 2 reports the weights on the 12 disaggregates at three specific observations. It can be seen from Table 2 that generally all disaggregate components have a non-zero weight, although the weight on Clothing and Footwear does drop to just below two percent.[16] There does not seem to be

---

[15]Smith (2004) and Kiley (2008) discuss the point forecasting properties of various core inflation measures. Most fail to outperform simple AR benchmarks.

[16]Geweke (2009) argues that even a zero weight is not sufficient to conclude that a component model

Table 2: Disaggregate weights

	1996Q3	2002Q3	2008Q4
Food and non-alc. bev.	0.138	0.128	0.117
Alc. bev. and tobacco	0.032	0.041	0.050
Cloth. and footwear	0.031	0.017	0.016
Housing, water, el. and fuel	0.105	0.091	0.070
Furnishings and house equip.	0.155	0.123	0.110
Health care	0.043	0.054	0.059
Transport	0.048	0.072	0.082
Communications	0.021	0.026	0.032
Recreation	0.136	0.161	0.159
Education	0.079	0.057	0.062
Rest. and hotels	0.139	0.139	0.135
Miscellaneous goods and services	0.071	0.090	0.107

Note: The columns reports disaggregate weights in three observations, 1996Q3, 2002Q3 and 2008Q4.

a case for excluding the information on individual disaggregates, or groups of particular disaggregates, on the basis of these weights.[17]

In figure 1, we plot the median from our DE12 density forecast, together with the 25 and 75 percentiles from this ensemble density. The plot shows that the median of the DE12 core ignores several extreme values in the actual measured inflation series. Typically, the probability of inflation being less than zero is well below 25 percent.
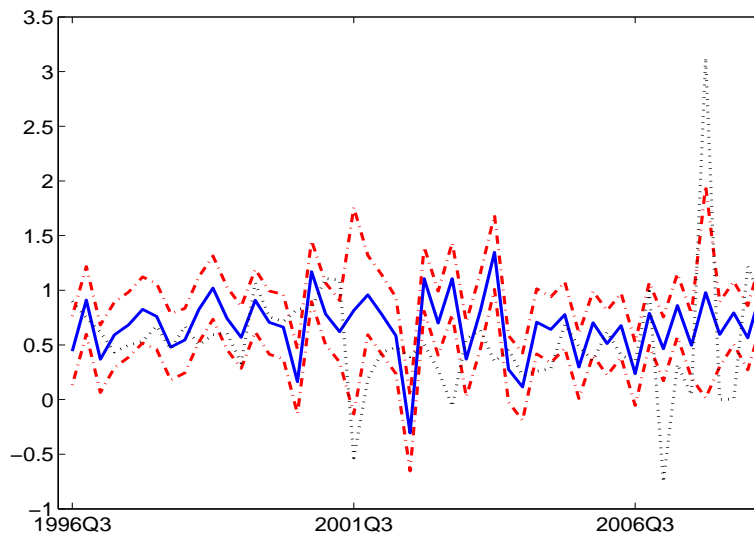
To provide further insight into the probability of tail events for inflation, figure 2 provides the ensemble predictive densities at particular observations, namely 1996Q3 and 2008Q4, the first and the last values in our evaluation period. We see that the AR(4) benchmark produces density forecasts that are too wide, with a high probability mass attributed to (quarterly) inflation of greater than two percent in absolute value for both observations. The DE12 predictives contain more mass in the regions around the outturn than the AR(4) benchmark, with relatively minor departures from symmetry.

We conclude from this analysis that the ensemble approach provides a means of generating accurate forecast densities for measured inflation from disaggregate information. Moreover, the common practice of discarding disaggregate information, either by zero-weighting groups or individual disaggregates as in the core inflation literature, is unwar-

---

has zero value for the linear opinion pool.

[17]Rules of thumb for truncation, such as, dropping disaggregates with less than 10 percent weight, may result in improvements in forecast performance. One difficulty to be explored in subsequent research is how to deal with the uncertainty over the truncation factor (eg 10 percent).

Figure 1: Inflation interval forecasts

Note: The figure shows the posterior median (solid line) of the predictive density given by disaggregate ensemble DE12 and actual inflation (the dashed line). The red dashed lines in the graphs are the 25th and 75th percentiles of the predictive density.
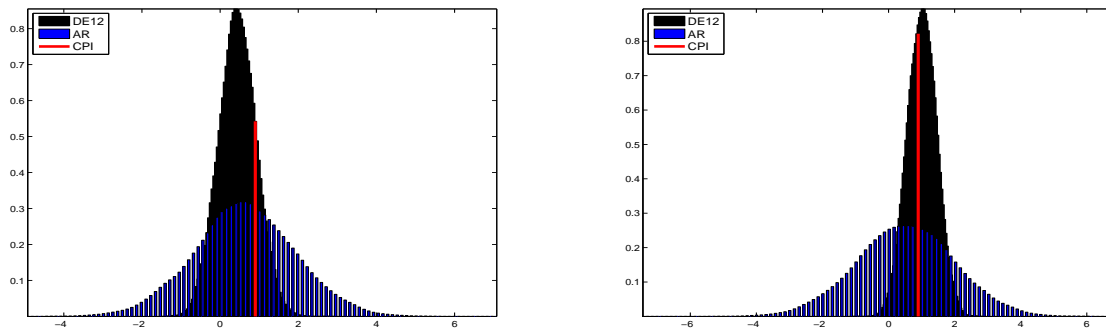
ranted in this density forecasting context as they do contain useful information about the future values of inflation.

3.3   A DSGE ensemble

Monetary policymakers typically use DSGE models as core workhorse models for forecasting and policy analysis. The introduction of Inflation Targeting led central banks to focus their macro models on inflation issues. One issue in doing so is that the current generation of models are considerably more abstract than the large-scale Keynesian macro models of the 1970s. Also, the DSGE tradition adopted by Inflation Targeters takes optimizing behaviour by micro agents as the cornerstone of model building. Despite the profusion of nominal and real rigidities adopted in the workhorse models, many critics argue that the models are fundamentally misspecified. We interpret this view as consistent with the incomplete model space concept in Geweke (2009).

In this example, we construct ensemble predictives for inflation based on the out of sample forecast performance of many component models, where all component models use a particular DSGE specification. The components are distinguished by the assumed

13

Figure 2: AR and DE12 density forecasts

(a) 1996Q3-2008Q4

Note: The figures plot the histogram of the density forecasts given by AR benchmark and by the disaggregate ensemble DE12 for two different periods, the first and last forecasts. The realized value for CPI is also provided.

(single candidate) break date in the sample, with each component using a unique post-break sample to produce forecasts through the evaluation period. The recursive simulation strategy for the DSGE model in this example follows closely the approach of Bache, Jore, Mitchell and Vahey (2009). Using Norwegian data, they compare the out of sample forecasting performance of NEMO (the Norges Bank core macro model) with benchmark models. In our paper, every component model is a replicant of NEMO, but with different start dates for in-sample estimation. We refer to the ensemble of DSGEs as EDSGE, which we construct by using CRPS weights for measured inflation. That is, we treat the component models in exactly the same way as in the previous example, and eschew multivariate density scoring.[18]

We emphasize that our EDSGE framework uses many very similar DSGE models. In each component, the agents and the government are assumed to have rational expectations. There is no learning taking place in any of our component models. In contrast, Svensson and Williams (2007) consider Markov jump-linear-quadratic systems that nest several prevalent, but relatively simple, macro models. It remains to be seen whether the Svensson-Williams approach will yield accurate forecast densities.

We demonstrate below that the EDSGE predictives provide accurate forecast densities for measured inflation using the same metrics for performance as in the previous (disaggregate) example. Simply eyeballing the weights illustrates that all of these components are quite likely, based on post-data analysis of density fit. The weights are pretty much

---

[18]Multivariate extensions and loss-based weighting of components pose no particular conceptual issues.

14

uniformly distributed across the components.

Since we have already described how the EDSGE will be constructed, we simply summarize the DSGE model, before turning to the structure of the components and then the results.

NEMO is a medium-scale New Keynesian small open economy model with a similar structure to the DSGE models recently developed in many other central banks. In this example, we use a simplified version of the model motivated by the need to reduce the computational burden of producing the recursive forecasts for forecast density combination. The simplification involves modifications to the Bayesian simulation methodology and the steady-state behaviour of the model as described below.

An appendix to Bache, Jore, Mitchell and Vahey (2009) describes the NEMO economy in detail. Here we summarize the main features. There are two production sectors. Firms in the intermediate goods sector produce differentiated goods for sale in monopolistically competitive markets at home and abroad, using labour and capital as inputs. Firms in the perfectly competitive final goods sector combine domestically produced and imported intermediate goods into an aggregate good that can be used for private consumption, private investment and government spending. The household sector consists of a continuum of infinitely-lived households that consume the final good, work and save in domestic and foreign bonds. The model incorporates real rigidities in the form of habit persistence in consumption, variable capacity utilization of capital and investment adjustment costs, and nominal rigidities in the form of local currency price stickiness and nominal wage stickiness. The model is closed by assuming that domestic households pay a debt-elastic premium on the foreign interest rate when investing in foreign bonds. A permanent technology shock determines the balanced growth path. The fiscal authority runs a balanced budget each period; and, the central bank sets the short-term nominal interest rate according to a simple monetary policy rule. The exogenous foreign variables are assumed to follow autoregressive processes. To solve the model we first transform the model into a stationary representation by detrending by the permanent technology shock. We then take a first-order approximation (in logs) of the equilibrium conditions around the steady-state.

Estimation uses data on the following ten variables: GDP, private consumption, business investment, exports, the real wage, the real exchange rate, overall inflation, imported inflation, the 3-month nominal money market rate, and hours worked. We measure inflation with the (headline) consumer price index adjusted for tax and energy prices—known as the 'CPIATE' measure. The interest rate is the 3-month money market rate, and the

Table 3: Forecast performance

	LR	LS	LS-test
AR	0.101	-0.657	
EDSGE	0.226	-0.125	0.020

Note: The column LR is the Likelihood Ratio p-value of the test of zero mean, unit variance and independence of the inverse normal cumulative distribution function transformed *PITS*, with a maintained assumption of normality for transformed *PITS*. LS is the average logarithmic score, averaged over the evaluation period. LS-test is the p-value of the KLIC-based test for equal density forecasting performance of the AR model and the EDSGE over the sample 2000Q1 to 20008Q4.

(seasonally adjusted) GDP variable excludes the oil and gas sectors.[19] Since the model predicts that domestic GDP, consumption, investment, exports and the real wage are non-stationary, these variables are included in first differences. We take the log of the real exchange rate and hours worked. All variables are demeaned prior to estimation.

We estimate the structural parameters using Bayesian techniques.[20] The structural parameters are re-estimated in each recursion for the evaluation period. We construct the forecast densities by drawing 10,000 times from a multivariate normal distribution for the shocks. The standard deviations of the shocks are set equal to their estimated posterior mode. Note that the (implicit) steady-states vary by recursion through the evaluation period; we demean the data prior to estimation in each recursion. We emphasize that, as a result of this simulation approach, our components do not account for parameter uncertainty and that the resulting predictives from each component are Gaussian.

We work with 14 component DSGE models, distinguished by the assumed start date for in-sample estimations. The longest sample used starts in 1985Q2; the last sample starts in 1988Q3. The other variants explore every feasible start date between. Estimation is based on expanding window samples. The evaluation period for the predictives is 2000Q1 to 2008Q4; the period 1999Q1 to 2000Q1 we use as a 'training period' to initialize the ensemble weights. This application focuses entirely on one-step ahead forecasts.

---

[19]The national accounts data relate to the mainland economy, that is, the total economy excluding the petroleum sector. See table 5 for details about the data and the sources.

[20]We carry out DSGE estimation in DYNARE; see Juillard (1996) and the following website http://www.cepremap.cnrs.fr/juillard/dynare/.

Table 4: DSGE weights

	2000Q1	2004Q4	2008Q4
DSGE-1985Q2	0.076	0.072	0.072
DSGE-1985Q3	0.077	0.071	0.072
DSGE-1985Q4	0.072	0.074	0.072
DSGE-1986Q1	0.073	0.077	0.072
DSGE-1986Q2	0.070	0.073	0.070
DSGE-1986Q3	0.061	0.069	0.067
DSGE-1986Q4	0.070	0.073	0.070
DSGE-1987Q1	0.083	0.073	0.073
DSGE-1987Q2	0.083	0.071	0.074
DSGE-1987Q3	0.070	0.067	0.070
DSGE-1987Q4	0.061	0.071	0.073
DSGE-1988Q1	0.067	0.070	0.072
DSGE-1988Q2	0.067	0.070	0.072
DSGE-1988Q3	0.069	0.068	0.071

Note: The columns report the weights, in three specific time periods, on the 13 components, differentiated by the proposed break-date, in the EDSGE. Each DSGE component is labeled by its break date.
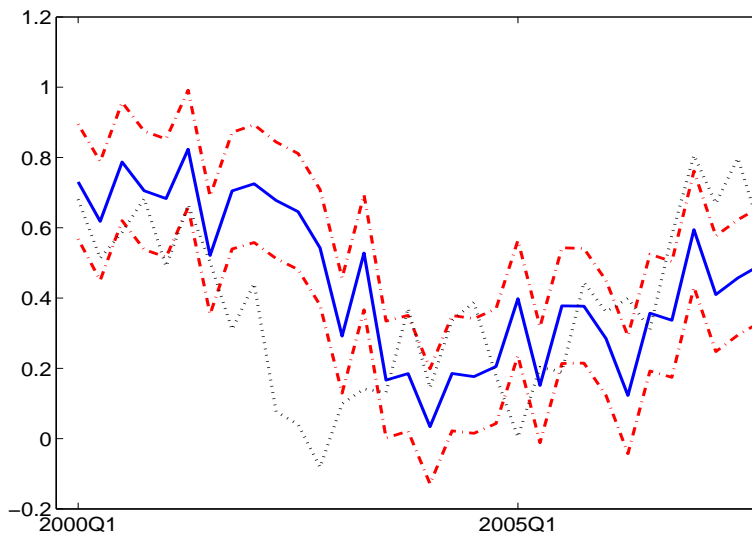
Before turning to the density evaluations for our EDSGE, we consider the performance of the point forecasts. The root mean squared prediction error (RMSPE) of EDSGE and the benchmark AR(4) is 0.074 and 0.027, respectively. That is, unlike the disaggregate ensemble in the previous example, our DSGE ensemble does not beat an autoregressive benchmark. This property stems from some mean bias in the components—none of the component models outperform the benchmark either.[21]

We turn now to ex post (end of evaluation period) evaluation of forecast densities from the EDSGE and the AR(4) benchmark. Table 3 has two rows which refer to the EDSGE and the AR benchmark. The columns report the Berkowitz likelihood ratio test (based on the *PITS*), the log scores (averaged over the evaluation period), and the *p*-values for the equal predictive density accuracy test (based on the log scores), respectively. Whereas both models appear well-calibrated on the basis of the Berkowitz likelihood ratio, the final column shows that the AR is rejected in favour of EDSGE using the KLIC-based test. EDSGE delivers a statistically significant, at the 98% level, improvement in the log score (reported in the second column).

We see in Table 4 that the weights on the different components in EDSGE display

---

[21]There may be scope to remove the forecast bias prior to combination. See, for example, Bao *et al* (2009).

Figure 3: Inflation interval forecasts

Note: The figure shows the posterior median (solid line) of the predictive density given by disaggregate ensemble EDSGE and the actual inflation (in dashed line). The red dashed lines in the graphs are the 25th and 75th percentiles of the predictive density.
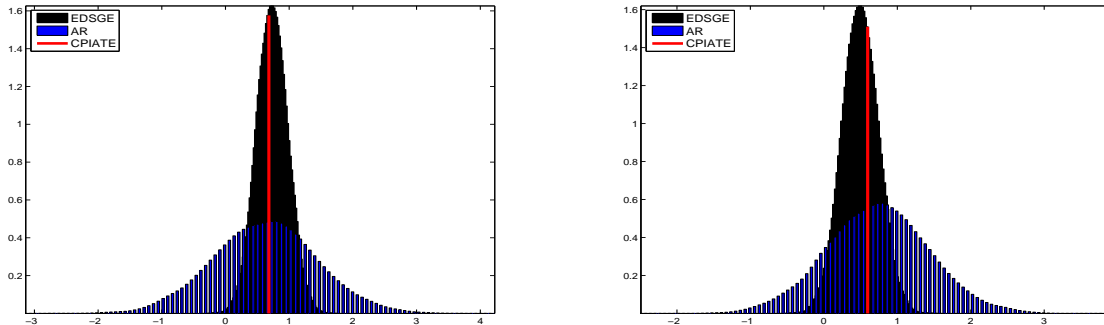
little volatility over time. And typically, all the DSGE components receive similar weight, which gives an indication of the individual plausibility of the components. The data suggest that a single DSGE model (a single break-date) should not be used for density forecasting.

In figure 3, we plot the median from our EDSGE density forecast, together with the 25 and 75 percentiles from the ensemble density. The plot shows that the median of the EDSGE is typically less volatile than the actual inflation series. Also apparent from this plot is the tendency for the runs of observations to occur outside the displayed percentiles. Clearly, there is scope for improving calibration by considering other candidate uncertainties—which could then be integrated out using the methods described in this paper. We leave this avenue to be explored in a more complete analysis of DSGE ensembles.

Figure 4 provides the ensemble predictive densities at particular observations, namely 2000Q1 and 2008Q4. We see that the EDSGE density is much sharper than the AR benchmark, and there are also some minor departures from symmetry.

We conclude from this analysis that the ensemble approach provides accurate forecast

18

Figure 4: AR and EDSGE density forecasts

(a) 2000Q1-2008Q4

Note: The figures plot the histogram of the density forecasts given by AR benchmark and by the ensemble EDSGE for two different periods, the first and last forecasts. The realized value for CPIATE is also provided.

densities for measured inflation based on the DSGE components. The weights on the components indicate both that none of the DSGE components are implausible and also that one component alone is not preferred over the others.

4   Conclusions and ideas for further research

We have argued that the next generation of macro modellers at Inflation Targeting central banks should adapt a methodology from the weather forecasting literature known as 'ensemble modelling'. In this approach, uncertainty about model specifications (e.g., initial conditions, parameters, and boundary conditions) is explicitly accounted for by constructing ensemble predictive densities from a large number of component models. The components allow the modeller to explore a wide range of uncertainties; and the resulting ensemble 'integrates out' these uncertainties using time-varying weights on the components. We have provided two specific examples of this modelling strategy.

The next generation of macro models at Inflation Targeting central banks could set aside the 'uncertain instabilities' problem and focus on these issues from the perspective of a single model. But a more promising route, we feel, is to explore model uncertainty explicitly. The computational simplicity of ensemble systems makes them very convenient for combining the evidence in many, highly-complex DSGE models, such as NEMO. Moreover, as our second example has demonstrated, the ensemble approach has the potential to produce accurate forecast densities from models used by policymakers in practice.

19

In discussing this paper at the Norges Bank Inflation Targeting conference, a number of participants argued that modelling expectations, learning and monetary policy strategy would pose serious challenges within the ensemble framework. We agree. But we think it a worthwhile endeavour.

Table 5: Variable definitions and sources. Observable variables in estimation of DSGE model.

Observables	Description	Source
Y_t	GDP mainland Norway, per capita, s.a.	Statistics Norway
C_t	Private consumption, per capita, s.a.	Statistics Norway
I_t	Business investment, per capita, s.a.	Statistics Norway
M_t^*	Exports mainland Norway, per capita, s.a.	Statistics Norway
W_t/P_t	Hourly wage income divided by private consumption deflator, s.a.	Statistics Norway
RER_t	Import-weighted real exchange rate (I-44)	Norges Bank
P_t	Overall price level adjusted for taxes and excl. energy prices (CPI-ATE), s.a.	Statistics Norway
P_t^m	Imported consumer prices adjusted for taxes and excl. energy prices, s.a.	Statistics Norway
R_t	3-month money market rate (NIBOR)	Norges Bank
l_t	Total hours worked, per capita, s.a.	Statistics Norway

References

[1] Andersson, M. and Karlsson, S. (2007), "Bayesian Forecast Combination for VAR Models", Sveriges Riksbank Working Paper Series No. 216.

[2] Amisano, G. and R. Giacomini (2007), "Comparing Density Forecasts via Likelihood Ratio Tests", *Journal of Business and Economic Statistics*, 25, 2, 177-190.

[3] Amisano, G. and J. Geweke (2009), "Evaluating the Predictive Distributions of Bayesian Models of Asset Returns", *International Journal of Forecasting* (forthcoming). Available as Working Paper 969, European Central Bank.

[4] Bache, I.W., A.S. Jore, J. Mitchell and S.P. Vahey (2009), "Combining VAR and DSGE Forecast Densities", Norges Bank, mimeo.

[5] Bao, Y., T-H. Lee and B. Saltoglu (2007), "Comparing Density Forecast Models", *Journal of Forecasting*, 26, 203-225. First circulated as 'A test for density forecast comparison with applications to risk management', University of California, Riverside, 2004.

[6] Bao, L., T. Gneiting, E.P. Grimit, P. Guttop, and A.E. Raftery (2007) "Bias Correction and Bayesian Model Averaging for Ensemble Forecasts of Surface Wind Direction", University of Washington, Department of Statistics, Technical Report, No 557.

[7] Berkowitz, J. (2001), "Testing Density Forecasts, with Applications to Risk Management", *Journal of Business and Economic Statistics*, 19, 465-474.

[8] Brischetto, A. and A. Richards (2006), "The Performance of Trimmed Mean Measures of Underlying Inflation", RBA Research Discussion Papers 2006-10, Reserve Bank of Australia.

[9] Clark, T.E. and M. W. McCracken (2009) "Averaging Forecasts from VARs with Uncertain Instabilities", *Journal of Applied Econometrics* (forthcoming). Available as Federal Reserve Bank of Kansas City Working Paper 06-12.

[10] Clark, T.E. and K.D. West (2006), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models", *Journal of Econometrics*, 138, 291-311.

[11] Clements, M. P. (2004), "Evaluating the Bank of England Density Forecasts of Inflation", *Economic Journal,* 114, 855-877.

[12] Corradi, V. and N.R. Swanson (2006), "Predictive Density Evaluation", G. Elliot, C.W.J. Granger and A. Timmermann, eds, *Handbook of Economic Forecasting,* North-Holland, 197-284.

[13] Del Negro, M., Schorfheide, F., Smets, F. and Wouters, R. (2007), "On the Fit of New Keynesian Models", *Journal of Business and Economic Statistics*, 25(2), 143-162.

[14] Diebold, F.X., Gunther, T.A. and A.S. Tay (1998) "Evaluating Density Forecasts; with applications to financial risk management", *International Economic Review,* 39, 863-83.

[15] Durlauf S. and S.P. Vahey (2009) "Editorial: Model Uncertainty and Macroeconomics", *Journal of Applied Eocnometrics*, forthcoming.

[16] Garratt, A., J. Mitchell and S.P. Vahey (2009), "Measuring Output Gap Uncertainty" unpublished manuscript, Birkbeck College, University of London.

[17] Gerard, H. and K. Nimark (2008), "Combining Multivariate Density Forecasts Using Predictive Criteria", RBA Research Discussion Papers 2008-02, Reserve Bank of Australia.

[18] Geweke, J. (2009), *Complete and Incomplete Econometric Models.* Princeton University Press (forthcoming).

[19] Giacomini, R. and H. White (2006), "Tests of Conditional Predictive Ability", *Econometrica*, 74, 1545-1578.

[20] Giordani, P. and R. Kohn (2008), "Efficient Bayesian Inference for Multiple Change-Point and Mixture Innovation Models", *Journal of Business and Economic Statistics*, 26, 66-77.

[21] Giordani, P., R. Kohn and D. van Dijk (2007), "A Unified Approach to Nonlinearity, Structural Change, and Outliers", *Journal of Econometrics*, 137, 112-133.

[22] Gneiting, T. and A. Raftery (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation", *Journal of the American Statistical Association*, 102, 359-378.

[23] Goodhart, C. (2007), "Whatever became of the monetary aggregates?", *National Institute Economic Review,* 200, 56-61.

[24] Hall, S.G. and J. Mitchell (2007), "Density Forecast Combination", *International Journal of Forecasting,* 23, 1-13.

[25] Jore, A. S., J. Mitchell and S.P. Vahey (2008), "Combining Forecast Densities from VARs with Uncertain Instabilities", *Journal of Applied Econometrics* (forthcoming). Available as NIESR Discussion Paper No. 303.

[26] Juillard, M. (1996), "DYNARE: A Program for the Resolution and Simulation of Dynamic Models with Forward Variables Through the use of a Relaxation Algorithm", CEPREMAP, Couverture Orange, 9602.

[27] Karagedikli, O., Matheson, T., Smith C. and S. P. Vahey, (2009), "RBCs and DSGEs: The Computational Approach to Business Cycle Theory and Evidence", *Journal of Economic Surveys* (forthcoming). Available as Reserve Bank of New Zealand Discussion Paper 2007/15.

[28] Kascha, C. and F. Ravazzolo (2009), "Combining Inflation Density Forecasts", *Journal of Forecasting* (forthcoming). Available as Norges Bank Working Paper 2008/22.

[29] Kiley, M. (2008), "Estimating the Common Trend Rate of Inflation for Consumer Prices and Consumer Prices Excluding Food and Energy Prices", Federal Reserve Board, FEDS 2008-38, July.

[30] Koop, G. (2003), *Bayesian Econometrics.* John Wiley and Sons.

[31] Leutbecher, M. and T.N. Palmer (1997), "Ensemble Forecasting", *Journal of Computational Physics*, 227, 3515-3539.

[32] MacKenzie, D. (2003), "Ensemble Kalman Filters Bring Weather Models Up to Date", *SIAM News*, 36, 8, October.

[33] Maheu, J.M. and S. Gordon (2008), "Learning, Forecasting and Structural Breaks", *Journal of Applied Econometrics*, 23, 553-583.

[34] Maheu, J.M. and T.H. McCurdy (2009), "How Useful are Historical Data for Forecasting the Long-Run Equity Return Distribution?", *Journal of Business and Economic Statistics*, 27, 95-112.

[35] Mandell, J. (2007) "A Brief Tutorial on the Ensemble Kalman Filter", Center for Computational Mathematics Reports No 242, University of Colarado at Denver.

[36] Mitchell, J. and S.G. Hall (2005), "Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR "Fan" Charts of Inflation", *Oxford Bulletin of Economics and Statistics*, 67, 995-1033.

[37] Mitchell, J. and K. F. Wallis (2009), "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness", NIESR Discussion Paper No. 320 and Warwick University Discussion Paper.

[38] Molteni, F., Buizza, R., Palmer, T. N. and T. Petroliagis (1996), "The new ECMWF ensemble prediction system: methodology and validation", *Quarterly Journal of the Royal Meteorological Society*, 122, 73-119.

[39] Patton, A. (2004), "On the Out-of-Sample Importance of Skewness and Asymmetric Dependence for Asset Allocation", *Journal of Financial Econometrics*, 2, 130-168.

[40] Panagiotelis, A. and M. Smith (2008), "Bayesian Density Forecasting of Intraday Electricity Prices using Multivariate Skew t Distributions", *International Journal of Forecasting*, 24, 710-727.

[41] Ravazzolo, F. and S.P. Vahey (2009), "Ditch the Ex! Measure Core Inflation with a Disaggregate Ensemble", Norges Bank unpublished manuscript.

[42] Sims, C. (2007) "Comment on Del Negro, Schorfheide, Smets and Wouters", *Journal of Business and Economic Statistics*, 25(2), 152-154.

[43] Smith, J.K. (2004) "Weighted Median Inflation: is this core inflation?", *Journal of Money, Credit and Banking*, 36, 253-263.

[44] Smith, C. et al. (2009) "More than One Weight to Skin a Cat: combining densities at Norges Bank", Norges Bank unpublished manuscript.

[45] Stock, J.H. and M.W. Watson (2001), "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series", in R.F. Engle and H. White, eds., *Festschrift in Honour of Clive Granger* (Cambridge University Press, Cambridge) 1-44.

[46] Stock, J. H. & Watson, M.W. (2004), "Combination Forecasts of Output Growth in a Seven-country Data Set", *Journal of Forecasting*, 23, 405–430.

[47] Svensson, L. and N. Williams (2007), "Monetary Policy with Model Uncertainty: Distribution Forecast Targeting ", unpublished manuscript, Princeton University.

[48] Wallis, K.F (2005), "Combining Density and Interval Forecasts: a Modest Proposal", *Oxford Bulletin of Economics and Statistics*, 67, 983-994.