Combining VAR and DSGE Forecast Densities*

Ida Wolden Bache (Norges Bank)
Anne Sofie Jore (Norges Bank)
James Mitchell (NIESR)
Shaun P. Vahey (Melbourne Business School)

November 5, 2009

Abstract

A popular macroeconomic forecasting strategy takes combinations across many models to hedge against instabilities of unknown timing; see (among others) Stock and Watson (2004), Clark and McCracken (2010), and Jore et al. (2010). Existing studies of this forecasting strategy exclude Dynamic Stochastic General Equilibrium (DSGE) models, despite the widespread use of these models by monetary policymakers. In this paper, we combine inflation forecast densities utilizing an ensemble system comprising many Vector Autoregressions (VARs), and a policymaking DSGE model. The DSGE receives substantial weight (for short horizons) provided the VAR components exclude structural breaks. In this case, the inflation forecast densities exhibit calibration failure. Allowing for structural breaks in the VARs reduces the weight on the DSGE considerably, and produces well-calibrated forecast densities for inflation.

Keywords: Ensemble modeling; Forecast densities; Forecast evaluation; VAR models; DSGE models

JEL codes: C32; C53; E37

*The views represented in this paper are those of the authors and not Norges Bank. We are grateful to Jon Nicolaisen for helpful conversations that led to this study. We also benefited greatly from comments by Mu-Chun Wang and participants at the Forecasting and Monetary Policy Conference (Bundesbank, Berlin, March 2009) and the 5th Dynare Conference (Norges Bank, Oslo, August-September 2009). We thank the ESRC (RES-062-23-1753) and the ARC (LP 0991098) for financial support.

1 Introduction

A common macroeconomic forecasting strategy combines the forecasts from many models to hedge against instabilities of unknown timing. Recent studies include (among others) Stock and Watson (2004), Clark and McCracken (2010), and Jore et al. (2010). The first two papers take the view that equal-weighting of component models can produce good point forecasts; the last study provides an example in which components weighted by the logarithmic score produce well-calibrated ensemble forecast densities.

The existing macro forecast combination literature excludes Dynamic Stochastic General Equilibrium (DSGE) models, in part because of the computational burden imposed by DSGE forecasting. Their absence presents a practical difficulty for the implementation of forecast combination methods at central banks. Many central banks have invested heavily in DSGE models, in which the growth model yields a linear and constant parameter reduced-form representation. The point forecasts from this class of (constant parameter) policymaking DSGE model typically match the accuracy of benchmark autoregressive representations.

In this paper, we use an expert combination framework to produce ensemble forecast densities for inflation from a system of Vector Autoregressions (VARs) and a DSGE model. The forecasts from the models are combined using the logarithmic score of the VAR forecast densities; see (among others) Jore et al. (2010). We evaluate the VAR-DSGE ensemble forecast densities for inflation using probability integral transforms (*pits*). This offers a means of evaluating density forecasts for general but unknown loss functions. We compare and contrast the calibration properties of the VAR-DSGE ensemble with and without structural break VAR components.

To ensure relevance for policymakers, we selected a DSGE model used routinely in practice by an Inflation Targeting central bank; namely NEMO, the Norges Bank core policymaking model. This DSGE model shares many features typical of the DSGE class of policymaking model. As one might expect given the conventional wisdom in the DSGE literature, the properties of the point forecasts for inflation from this model match the accuracy of benchmark autoregressive specifications.

Turning to our results, we find that our VAR-DSGE ensemble which allows for breaks gives well-calibrated forecast densities for inflation, but relatively small weight is attached to the DSGE. Restricting attention to VAR components without breaks, the DSGE receives a substantial weight for some horizons. For example, for the one-step ahead case, the weight on the DSGE model reaches 60 percent by the end of our evaluation. But without break components the VAR-DSGE ensemble produces poorly calibrated forecast densities at the horizons for which the DSGE receives most weight. The break variants capture the shifts in volatility present in our "Great Moderation" sample. Although unimportant for point forecast evaluations based on Root Mean Squared Forecast Error (RMSFE), excluding break components impairs forecast density calibration; see Clark (2009) for a discussion of parameter change and VAR forecast density calibration.

The plan of the paper is as follows. In the subsequent section, we outline our methods for ensemble forecasting. In Section 3, we describe our component models and data; and in Section 4, we summarize our recursive forecasting exercise. Our results are presented in Section 5. Some ideas for further research are contained in the final section.

2 Methods for ensemble forecasting

We construct the predictive densities from the large number of component models using forecast density combination methods. Earlier papers by Jore et al. (2010) and Garratt et al. (2009) take this approach to ensemble modeling. Although point forecast combination has a longer tradition in economics (e.g., see Bates and Granger (1969)), the focus of this study is on providing monetary policymakers with an estimate of the entire probability distribution of the possible future values of the variable of interest—the forecast density. Many Inflation Targeting central banks (including the Bank of England, Norges Bank, and Sveriges Riksbank) provide forecast densities for inflation (and other variables) to communicate the policy stance.[1]

2.1 Forecast density combination

The ensemble densities are constructed using an expert combination methodology, utilizing the linear opinion pool; see, for example, Morris (1974, 1977), Winkler (1981), Lindley (1983) and Genest and McConway (1990):

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

where $g(X_{\tau,h} \mid I_{i,\tau})$ are the h-step ahead forecast densities from component model i, $i = 1, \ldots, N$ of a random variable X_τ, (with realization x_τ), conditional on the information

[1] The Federal Reserve recently moved to publish FOMC member forecasts—in effect, a forecast interval.

set I_τ. The non-negative weights, $w_{i,\tau,h}$, in this finite mixture sum to unity.[2] Furthermore, the weights may change with each recursion in the evaluation period $\tau = \underline{\tau}, \dots, \overline{\tau}$. Since the ensemble density defined by equation (1) is a mixture it delivers a more flexible distribution than each of the component densities from which it is derived. As N increases, the ensemble density becomes more and more flexible, with the potential to approximate non-linear specifications.

We construct the ensemble weights based on the fit of the individual model forecast densities. Like Amisano and Giacomini (2007) and Hall and Mitchell (2007), we use the logarithmic score to measure density fit for each component model through the evaluation period. The logarithmic score of the i-th density forecast, $\ln g(X_{\tau,h} \mid I_{i,\tau})$, is the logarithm of the probability density function $g(. \mid I_{i,\tau})$, evaluated at the outturn $x_{\tau,h}$. The logarithmic scoring rule gives a high score to a density forecast that assigns a high probability to the realized value. Following Jore et al. (2010) the recursive weights for the h-step ahead densities take the form:

$$w_{i,\tau,h} = \frac{\exp\left[\sum_{\underline{\tau}-10}^{\tau-h} \ln g(x_{\tau,h} \mid I_{i,\tau})\right]}{\sum_{i=1}^{N} \exp\left[\sum_{\underline{\tau}-10}^{\tau-h} \ln g(x_{\tau,h} \mid I_{i,\tau})\right]}, \qquad \tau = \underline{\tau}, \dots, \overline{\tau} \tag{2}$$

where $\underline{\tau}-10$ to $\underline{\tau}$ comprises the training period used to initialize the weights. Computation of these weights is feasible for a large N ensemble. Given the uncertain instabilities problem, the recursive weights should be expected to vary across τ.

Our ensemble methodology has many similarities with an approximate predictive likelihood approach (for the one-step horizon); see Raftery and Zheng (2003), and Eklund and Karlsson (2007). Given our definition of density fit, the model densities are combined using Bayes' rule with equal (prior) weight on each model—which a Bayesian would term non-informative priors. Garratt et al. (2009) discuss ensemble modeling in macro-econometric applications, and other applied statistics fields. Bache et al. (2010) and Geweke (2009) discuss combinations in an incomplete model space in which the true model is likely absent.

2.2 Forecast density evaluations

A popular evaluation method for forecasts densities, following Rosenblatt (1952), Dawid (1984) and Diebold et al. (1998), evaluates relative to the "true" but unobserved density

[2] The restriction that each weight is positive could be relaxed; for discussion see Genest and Zidek (1986).

using the probability integral transforms (*pits*) of the realization of the variable with respect to the forecast densities. A density forecast can be considered optimal (regardless of the user's loss function) if the model for the density is correctly calibrated; i.e., if the *pits* $z_{\tau,h}$, where $z_{\tau,h} = \int_{-\infty}^{x_{\tau,h}} p(u)du$, are uniform and, for one-step ahead forecasts, independently and identically distributed. In practice, therefore, density evaluation with the *pits* requires application of tests for goodness-of-fit and independence at the end of the evaluation period.[3]

The goodness-of-fit tests employed include the Likelihood Ratio (LR) test proposed by Berkowitz (2001). Results are presented at $h > 1$ using a two degrees-of-freedom variant (without a test for autocorrelation, see Clements (2004)). For the one-step horizon, $h = 1$, we use a three degrees-of-freedom variant with a test for independence, where under the alternative $z_{\tau,h}$ follows an AR(1) process. Since the LR test has a maintained assumption of normality, we also consider the Anderson-Darling (AD) test for uniformity, a modification of the Kolmogorov-Smirnov test, intended to give more weight to the tails (and advocated by Noceti et al. (2003)). We also follow Wallis (2003) and employ a Pearson chi-squared test which divides the range of the $z_{\tau,h}$ into eight equiprobable classes and tests whether the resulting histogram is uniform. To test independence of the *pits*, we use a Ljung-Box (LB1) test, based on autocorrelation coefficients up to four.[4] For $h > 1$ we test for autocorrelation at lags greater than $(h - 1)$, but less than six, using a modified LB test (MLB). Even for correctly calibrated densities, we expect autocorrelation stemming from the overlapping forecast horizons.

3 Component models

Our ensemble forecast densities for inflation use a model space including many VARs and a (monetary policymaking) DSGE model. In this section, we describe the two types of model in detail.

[3]Given the large number of component densities under consideration, we do not allow for parameter uncertainty when evaluating the *pits*. Corradi and Swanson (2006) review *pits* tests computationally feasible for small N.

[4]To investigate possible higher order dependence we also undertook tests in the second and third powers of the *pits*. Results were similar to the first power.

3.1 VARs

We consider a range of VAR models in output growth, inflation and the interest rate. Table 3 in the Appendix describes the Norwegian data sources. Output growth refers to mainland economy GDP, seasonally adjusted (the total economy excluding the petroleum sector). We measure inflation with the (headline) consumer price index, adjusted for tax and energy prices, seasonally adjusted. The interest rate is the three month money rate (NIBOR).

The VARs are estimated with maximum lag lengths of 1 to 4. For each maximum lag order, we estimate trivariate VARs, bivariate VARs, and ARs, with inflation always included. This models space has 16 components in total: 4 ARs, 8 bivariate VARs, and 4 trivariate VARs. (For simplicity, we refer to these models collectively as VARs.) For each specification, we also transform the variables prior to estimation in two ways: we include first-differenced VARs (DVARs), and de-trended VARs, giving 48 models in total. Following Cogley (2002) and others, we use an exponential smoother to extract the trend, with the smoothing parameter set at 0.05; see also Clark and McCracken (2010).

We utilize a direct forecast methodology (see Marcellino et al. (2003)) to generate the h-step ahead predictive densities from each Gaussian Linear model. Given our non-informative priors, the predictive densities for the endogenous variables for each VAR component are multivariate Student-t; see Zellner (1971), pp. 233-236 and, for a more recent application, Garratt et al. (2009).

Following Jore et al. (2010) and Garratt et al. (2008), we add to the 48 specifications described above by adding variants with a single structural break in the conditional mean and variance, assuming a common break-date in all equations. That is, for each VAR specification, we estimate a separate VAR for every feasible start date for (in-sample) estimation from 1980Q1 to 1990Q4. With these additional structural break variants added to the 48 of full sample VARs, for each recursion in the evaluation period we consider 1777 component models in total. We shall construct ensembles both with and without structural breaks variants. We refer to these as “break” and “no break ” VAR ensembles.

3.2 The DSGE model: NEMO

NEMO is the core model used by Norges Bank for monetary policy. It 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, e.g., Sveriges Riksbank (see Adolfson et al. (2008)). In this paper, 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 simulation methodology and the steady-state behavior of the model as described below.

An appendix describes the NEMO economy in detail.[5] 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 labor 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 transform the model into a stationary representation, detrending by the permanent technology shock. Then, we take a first-order approximation (in logs) of the equilibrium conditions around the stochastic steady state.

Estimation uses data on the three variables used in the VAR specifications, plus the following seven variables: private consumption, business investment, exports, the real wage, the real exchange rate, imported inflation, and hours worked. The national accounts data relate to the mainland economy; that is, the total economy excluding the petroleum sector. Table 3 in the Appendix describes the data sources. 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 for each recursion.

We estimate the structural parameters using Bayesian techniques.[6] The forecast draws

[5] See Brubakk et al. (2006) for a more thorough discussion of NEMO.

[6] We carry out DSGE estimation in DYNARE; see Juillard (1996). Karagedikli et al. (2010) provide a simplified DSGE example with code.

are based on the mode of the posterior distributions for the structural parameters; the forecast densities for the DSGE model (like our VARs) do not allow for parameter uncertainty. 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. The values of the calibrated parameters and the priors for the estimated parameters are listed in table 4.

The sample used for estimation starts in $1987Q1$. This matches the practice used in Norges Bank monetary policymaking applications and differs from the starting point for the estimation of our full sample VARs, $1980Q1$. In effect, the (constant parameter) DSGE model uses a (slightly) shorter sample to restrict attention to the (Norwegian) Great Moderation period. Using a longer sample—to match the assumption in our VAR specifications without breaks—requires a larger variance for the shock processes to match the data. Bache et al. (2010) explore the density forecasting performance of this model with a range of starting points for estimation.

4 Recursive forecasting exercise

Recall that the DSGE model uses 10 observable for estimation and that the VARs include (up to) three variables. Hence, we conduct a "limited information" analysis of forecasting performance. Mindful of the Inflation Targeting regime adopted by many central banks, including Norges Bank, in presenting the results from our forecast analysis, we restrict our attention to inflation.[7]

Our recursive forecasting exercise is intended to mimic the behavior of a policymaker forecasting in real time. The information lags assumed are consistent with the release of the macro variables concerned. Unfortunately, we are not able to utilize real-time macroeconomic data because the data have not been compiled for (most of) the 10 observables used in DSGE estimation. Instead, we use a single vintage of data available in 2008Q4 for all forecasts and realizations.

The recursive forecast experiments are constructed as follows. We estimate each com-

[7]The qualitative results for output growth are similar to those for inflation. The interest rate forecast densities are poorly calibrated even for ensembles with components allowing for breaks, reflecting the difficulty of predicting interest rates with just three variables.

ponent on a sample ending in $\underline{\tau} - h$ and compute model forecasts for inflation for horizons of $h = 1, \ldots, 4$. We construct (recursive) ensemble predictive densities for $\underline{\tau} - h$ in the manner described in Section 2 using the weights given in equation (2). Then we extend the sample by one quarter, re-estimate each component, compute new h-step ahead forecasts for each model, and produce the ensemble forecast densities. This exercise is repeated over the evaluation period, $\underline{\tau} = 1998Q2$ to $\overline{\tau} = 2007Q3$.

5 Results

As a background to our forecast density evaluations, we remark briefly on the RMSFE of the two VAR ensembles (with and without breaks), and the DSGE. Earlier studies have drawn attention to the competitive point forecast performance of DSGEs; see, for example, Adolfson et al. (2008) and Schorfheide et al. (2009). Using a Diebold-Mariano test, we could not distinguish the forecasting performance of the break VAR ensemble, the no break VAR ensemble or the DSGE from an AR(1) benchmark at a 5% significance level for all forecast horizons, $h = 1, \ldots, 4$. We note, however, that the break ensemble did produce slightly lower RMSFE.

In Table 1 where there are four panels, one for each h, we turn to the *pits* tests on the forecast densities. To facilitate reading, we place the p-values in bold when the density forecast is correctly calibrated at a 5% significance level—that is, when we cannot reject the null hypothesis that the densities are correctly calibrated according to the evaluation test considered.[8] The first row of each panel in Table 1 shows that at shorter horizons ($h = 1$ and $h = 2$), the densities from the no break VAR ensemble do not appear to be well calibrated across the four tests. But at horizons, $h = 3$ and $h = 4$, we cannot reject the null of correct calibration in all cases.

The break VAR ensemble densities, evaluated in the second row of each panel, are well calibrated at 5% at all 4 horizons. Given our focus on performance of the ensemble, rather than its components, we do not report confidence intervals for the break dates. But suffice to say that the break models with the most support typically have a break-date prior to 1987Q1. Hence, the DSGE sample used for parameter estimation excludes the most likely, but not all, the candidate break dates (for the VAR model space).

Turning to the DSGE densities, evaluated in the third row of each panel of Table 1,

[8] To control the joint size of the four evaluation tests requires use of a stricter p-value. The Bonferroni correction suggests a p-value threshold, for an overall 5% signficance level, of $(100\% - 95\%)/4 = 1.25\%$ rather than 5% on each test. This gives qualitatively similar findings.

as with the no break VAR ensemble we see calibration failure for some individual tests at shorter horizons. The forecast densities are poorly calibrated according to three of the four *pits* tests. But, unlike the no break VAR ensemble, calibration is also weak at longer horizons. Note that, calibration failures occur despite the competitive point forecasting performance of the DSGE.

The preceding analysis has focused on forecast density evaluation for the break VAR ensemble, the no break VAR ensemble and the DSGE predictive densities. It is also instructive to combine the VAR and DSGE candidates to give a VAR-DSGE ensemble. Following Garratt et al. (2009), the combination of these predictive densities can be thought of as a "grand ensemble". In our application, one of the candidates is not an ensemble—the DSGE predictive density—but the density combination exercise is similar in spirit to that contained in Garratt et al. (2009).[9]

Figure 1 plots over time, at each of the four forecasting horizons, the weight on the DSGE in the VAR-DSGE ensemble in the two cases; that is, when the DSGE is combined with the break VAR ensemble and the no break VAR ensemble, respectively. Inspection of these grand ensemble weights indicates that the DSGE densities are competitive at shorter horizons, $h = 1$ and $h = 2$, against the no break VAR ensemble (solid line, top two panels). The DSGE density receives around 70 percent of the weight for $h = 1$ at the start of the evaluation, dropping to less than 20 percent in 2000, before returning non-monotonically to nearly its early level by the end of the evaluation. The DSGE weight for $h = 2$ starts at around 50 percent and declines throughout the evaluation (with some reversals) to roughly 10 percent.

In contrast, at the same short horizons, the weight on the DSGE in the grand ensemble with the break VARs is much lower (dashed line). In these cases, the DSGE receives around 20 percent weight at the start of the evaluation, with the contribution diminishing as the evaluation period increases. By the end of the evaluation period, the weight on the DSGE density is approximately zero at both $h = 1$ and $h = 2$.

At longer horizons (bottom two panels), $h = 3$ and $h = 4$, typically the DSGE density receives a smaller weight than at shorter horizon, regardless of whether the VAR ensemble includes break specifications or not. Even in the no break VAR case (solid line), the weight on the DSGE reaches roughly zero by the midpoint of the evaluation period. Recall that the no break VAR ensemble had a more satisfactory performance for longer horizons; see

[9] An extra training window is required to initialize the weights in the VAR-DSGE ensemble. We set this to $5 + h$ quarters for the first forecast, and use the training period information in the construction of the weights throughout our recursive forecasting exercise.

Table 1.

Table 2 contains the *pits* tests for the two VAR-DSGE grand ensembles. We note that both cases, based on combining the DSGE with the no break VAR ensemble (first row, for each horizon) or on combining the DSGE with the break VAR ensemble (second row, for each horizon) display good calibration at longer horizons. But, at shorter horizons, the no break VAR-DSGE ensemble displays calibration failure for one ($h = 1$) and two ($h = 2$) tests—excluding break VAR components matters least at longer horizons.

We emphasize that throughout our analysis we are using the recursive logarithmic score on the different specifications to construct the forecast densities. An alternative approach described by Hall and Mitchell (2007) and Geweke (2009) selects the combined predictive density with the highest average logarithmic score by iterative methods. Given the large number of models under consideration, this approach is infeasible for construction of our VAR ensembles. Nevertheless, we checked our findings for our VAR-DSGE grand ensembles (where there are only two specifications to be combined in each case), and found the weights to be broadly similar to those reported above.[10]

6 Conclusions

We draw the following conclusions from our evaluations of the forecast densities for inflation. First, ensemble densities based on component VARs with breaks are well calibrated. Second, ensembles from VARs without breaks exhibit poor calibration by some tests. Third, (despite competitive point forecast performance) the DSGE does not match the performance of break VAR ensembles. And finally, the DSGE receives a higher weight in the VAR-DSGE grand ensemble if the VAR ensemble does not contain components with breaks.

Our findings about the forecast density performance of a well-known policymaking DSGE model will spur further analysis. DSGE models which allow for time variation offer the scope for better predictive densities. Some recent candidates include Fernandez-Villaverde and Rubio-Ramirez (2007), and Justiniano and Primiceri (2008). To our knowledge, no central banks have yet adapted these methods for policy use, no doubt deterred by the computational burden of medium-sized DSGE models with time-varying parameter distributions. Our VAR-DSGE ensemble approach offers a computationally convenient methodology for producing well-calibrated forecast densities using a more conventional

[10] In two cases, at $h = 3$ and $h = 4$, the DSGE received slightly less weight than shown in Figure 1 throughout the evaluation.

log-linearized policymaking DSGE model. Although we emphasize that the break VAR ensemble performed well in our forecasting exercise without the benefit of the DSGE. In future work, we intend to investigate the suggestion in Bache et al. (2010) that ensemble combinations of many DSGE models can approximate non-Gaussian data generating processes. That research agenda offers the prospect of well-calibrated forecast densities from more structural components at a relatively low computational cost.

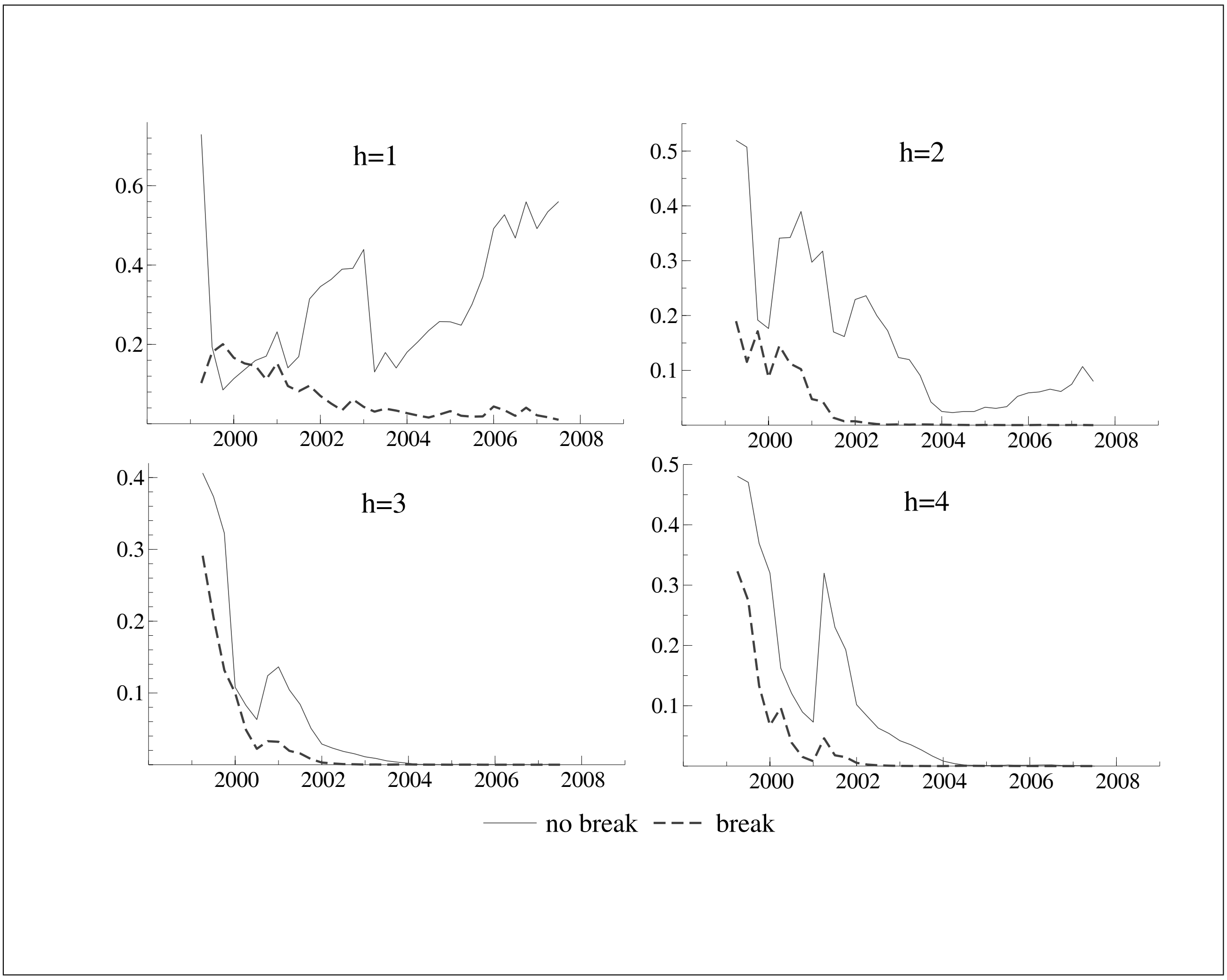

Figure 1: Weights on the DSGE in VAR-DSGE grand ensembles, by horizon h

Table 1: Density forecast evaluation of VAR ensemble and DSGE using *pits*

$h = 1$	LR3	AD	χ^2	LB1
no break VAR ensemble	0.03	**0.32**	**0.17**	**0.40**
break VAR ensemble	**0.38**	**0.49**	**0.17**	**0.54**
DSGE	0.00	0.04	**0.09**	0.02

$h = 2$	LR2	AD	χ^2	MLB
no break VAR ensemble	0.01	**0.16**	0.03	**0.51**
break VAR ensemble	**0.40**	**0.33**	**0.17**	**0.80**
DSGE	0.00	0.00	0.00	**0.99**

$h = 3$	LR2	AD	χ^2	MLB
no break VAR ensemble	**0.28**	**0.70**	**0.11**	**0.83**
break VAR ensemble	**0.32**	**0.25**	**0.59**	**0.71**
DSGE	0.00	0.00	0.00	**0.96**

$h = 4$	LR2	AD	χ^2	MLB
no break VAR ensemble	**0.45**	**0.85**	**0.71**	**0.15**
break VAR ensemble	**0.10**	**0.09**	**0.34**	**0.64**
DSGE	0.00	0.00	0.00	**0.97**

Notes: LR2 is the p-value for the Likelihood Ratio test of zero mean and unit variance of the inverse normal cumulative distribution function transformed pits, with a maintained assumption of normality for the transformed pits; LR3 supplements LR2 with a test for zero first order autocorrelation. AD is the p-value for the Anderson-Darling test statistic for uniformity of the pits, with the small-sample (simulated) p-values computed assuming independence of the pits. χ^2 is the p-value for the Pearson chi-squared test of uniformity of the pits histogram in eight equiprobable classes. LB is the p-value from a Ljung-Box test for independence of the pits; MLB is a modified LB test which tests for independence at lags greater than or equal to h

Table 2: Density forecast evaluation of the VAR-DSGE ensembles using *pits*

$h = 1$	LR3	AD	χ^2	LB1
no break VAR-DSGE	0.02	**0.27**	**0.29**	**0.48**
break VAR-DSGE	**0.30**	**0.42**	**0.26**	**0.46**

$h = 2$	LR2	AD	χ^2	MLB
no break VAR-DSGE	0.01	**0.10**	0.01	**0.64**
break VAR-DSGE	**0.35**	**0.28**	**0.54**	**0.75**

$h = 3$	LR2	AD	χ^2	MLB
no break VAR-DSGE	**0.26**	**0.66**	**0.26**	**0.29**
break VAR-DSGE	**0.27**	**0.22**	**0.65**	**0.64**

$h = 4$	LR2	AD	χ^2	MLB
no break VAR-DSGE	**0.25**	**0.56**	**0.95**	**0.21**
break VAR-DSGE	**0.10**	**0.07**	**0.22**	**0.57**

Notes: the no break VAR-DSGE refers to the grand ensemble of the no break VAR ensemble with the DSGE; the break VAR-DSGE refers to the grand ensemble of the break VAR ensemble with the DSGE. Also see notes to Table 1.

References

Adolfson, M., S. Lasen, J. Lind, and M. Villani (2008). Evaluating an estimated new keynesian small open economy model. *Journal of Economic Dynamics and Control 32*(8), 2690 – 2721.

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

Bache, I. W., J. Mitchell, F. Ravazzolo, and S. P. Vahey (2010). Macro modeling with many models. In D. Cobham, Ø. Eitrheim, S. Gerlach, and J. Qvigstad (Eds.), *Twenty Years of Inflation Targeting: Lessons Learned and Future Prospects*. Cambridge University Press. Forthcoming.

Bates, J. M. and C. W. J. Granger (1969). The combination of forecasts. *Operational Research Quarterly 20*, 451–468.

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

Brubakk, L., T. A. Husebø, J. Maih, K. Olsen, and M. Østnor (2006). Finding NEMO: Documentation of the norwegian economy model. Norges Bank Staff Memo 2006/6.

Clark, T. E. (2009). Real-time density forecasts from VARs with stochastic volatility. Research Working Paper RWP 09-08, Federal Reserve Bank of Kansas City.

Clark, T. E. and M. W. McCracken (2010). Averaging forecasts from VARs with uncertain instabilities. *Journal of Applied Econometrics*. Forthcoming. Revision of Federal Reserve Bank of Kansas City Working Paper 06-12.

Clements, M. P. (2004). Evaluating the Bank of England density forecasts of inflation. *Economic Journal 114*, 844–866.

Cogley, T. (2002). A simple adaptive measure of core inflation. *Journal of Money, Credit and Banking 34*(1), 94–113.

Corradi, V. and N. R. Swanson (2006). Predictive density evaluation. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, pp. 197–284. North Holland: North-Holland.

Dawid, A. P. (1984). Statistical theory: the prequential approach. *Journal of the Royal Statistical Society B 147*, 278–290.

Diebold, F. X., A. Gunther, and K. Tay (1998). Evaluating density forecasts with application to financial risk management. *International Economic Review 39*, 863–883.

Eklund, J. and S. Karlsson (2007). Forecast combination and model averaging using predictive measures. *Econometric Reviews 26*(2-4), 329–363.

Fernandez-Villaverde, J. and J. Rubio-Ramirez (2007). Estimating macroeconomic models: A likelihood approach. *Review of Economic Studies 74*(4), 1059–1087.

Garratt, A., G. Koop, E. Mise, and S. P. Vahey (2009). Real-time prediction with UK monetary aggregates in the presence of model uncertainty. *Journal of Business and Economic Statistics 27*(4), 480–491.

Garratt, A., G. Koop, and S. P. Vahey (2008). Forecasting substantial data revisions in the presence of model uncertainty. *Economic Journal 118*(530), 1128–1144.

Garratt, A., J. Mitchell, and S. P. Vahey (2009). Measuring output gap uncertainty. NIESR Discussion Paper No. 342.

Genest, C. and K. J. McConway (1990). Allocating the weights in the linear opinion pool. *Journal of Forecasting 9*, 53–73.

Genest, C. and J. Zidek (1986). Combining probability distributions: a critique and an annotated bibliography. *Statistical Science 1*, 114–135.

Geweke, J. (2009). *Complete and Incomplete Econometric Models.* Princeton University Press. Forthcoming.

Hall, S. G. and J. Mitchell (2007). Combining density forecasts. *International Journal of Forecasting 23*, 1–13.

Jore, A. S., J. Mitchell, and S. P. Vahey (2010). Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics*. Forthcoming. Available as NIESR Discussion Paper No. 303.

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.

Justiniano, A. and G. E. Primiceri (2008). The time-varying volatility of macroeconomic fluctuations. *American Economic Review 98*(3), 604–41.

Karagedikli, O., T. Matheson, C. Smith, and S. P. Vahey (2010). RBCs and DSGEs: The computational approach to business cycle theory and evidence. Reserve Bank of New Zealand Discussion Paper DP2007/15. Forthcoming. Journal of Economic Surveys.

Lindley, D. (1983). Reconciliation of probability distributions. *Operations Research 31*, 866–880.

Marcellino, M., J. Stock, and M. Watson (2003). A comparison of direct and iterated AR methods for forecasting macroeconomic series h-steps ahead. *Journal of Econometrics 135*, 499–526.

Morris, P. (1974). Decision analysis expert use. *Management Science 20*, 1233–1241.

Morris, P. (1977). Combining expert judgments: A Bayesian approach. *Management Science 23*, 679–693.

Noceti, P., J. Smith, and S. Hodges (2003). An evaluation of tests of distributional forecasts. *Journal of Forecasting 22*, 447–455.

Raftery, A. E. and Y. Zheng (2003). Long-run performance of Bayesian model averaging. *Journal of the American Statistical Association 98*, 931–938.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics 23*, 470–472.

Schorfheide, F., K. Sill, and M. Kryshko (2009). DSGE model-based forecasting of non-modelled variables. *International Journal of Forecasting*. Forthcoming. Available as Federal Reserve Bank of Philadelphia Working Paper 08-17.

Stock, J. H. and M. W. Watson (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting 23*, 405–430.

Wallis, K. F. (2003). Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts. *International Journal of Forecasting 19*, 165–175.

Winkler, R. (1981). Combining probability distributions from dependent information sources. *Management Science 27*, 479–488.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics.* New York: John Wiley and Sons.

Appendix: Structure of NEMO DSGE model

Final goods sector The perfectly competitive final goods sector consists of a continuum of final good producers indexed by $x \in [0,1]$ that aggregates domestic intermediate goods, Q, and imports, M, using a CES technology:

$$A_t(x) = \left[\eta^{\frac{1}{\mu}} Q_t(x)^{1-\frac{1}{\mu}} + (1-\eta)^{\frac{1}{\mu}} M_t(x)^{1-\frac{1}{\mu}}\right]^{\frac{\mu}{\mu-1}}, \tag{3}$$

The degree of substitutability between the composite domestic and imported goods is determined by the parameter $\mu > 0$, whereas η $(0 \leq \eta \leq 1)$ measures the steady-state share of domestic intermediates in the final good for the case where relative prices are equal to 1. The composite good $Q(x)$ is an index of differentiated domestic intermediate goods, produced by a continuum of firms $h \in [0,1]$:

$$Q_t(x) = \left[\int_0^1 Q_t(h,x)^{1-\frac{1}{\theta_t}}\, dh\right]^{\frac{\theta_t}{\theta_t - 1}}, \tag{4}$$

where the degree of substitution between domestic intermediate goods, θ_t, evolves according to AR(1) process with autoregressive parameter λ^{θ} and standard deviation σ_{θ}. Similarly, the composite imported good is a CES aggregate of differentiated import goods indexed $f \in [0,1]$:

$$M_t(x) = \left[\tfrac{1}{\theta^*} \int_0^1 M_t(f,x)^{1-\frac{1}{\theta^*}}\, df\right]^{\frac{\theta^*}{\theta^* - 1}}, \tag{5}$$

where θ^* is the degree of substitution between imported goods.

Intermediate goods sector Each intermediate firm h is assumed to produce a differentiated good $T_t(h)$ for sale in domestic and foreign markets using a CES production function:

$$T_t(h) = \left[(1-\alpha)^{\frac{1}{\xi}} \left(Z_t z_t^L l_t(h)\right)^{1-\frac{1}{\xi}} + \alpha^{\frac{1}{\xi}} \overline{K}_t(h)^{1-\frac{1}{\xi}}\right]^{\frac{\xi}{\xi-1}}, \tag{6}$$

where $\alpha \in [0,1]$ is the capital share and ξ denotes the elasticity of substitution between labor and capital. The variables $l_t(h)$ and $\overline{K}_t(h)$ denote, respectively, hours used and effective capital of firm h in period t. There are two exogenous shocks to productivity in the model: Z_t refers to an exogenous permanent (level) technology process, which grows at the gross rate π_t^z and evolves according to an AR(1) process with autoregressive parameter λ^z and standard deviation σ_z, whereas z_t^L denotes a temporary (stationary) shock to productivity (or labor utilization) which evolves according to an AR(1) process with autoregressive parameter λ^L and standard deviation σ_L. The variable $K_t(h)$ is defined as firm h's capital stock that is chosen in period t and becomes productive in period $t+1$. Firm h's *effective* capital in period t is related to the capital stock that was chosen in period $t-1$ by

$$\overline{K}_t(h) = u_t(h) K_{t-1}(h), \tag{7}$$

where $u_t(h)$ is the endogenous rate of capital utilization. Adjusting the utilization incurs a cost of $\gamma_t^u(h)$ units of final goods per unit of capital. The cost function is

$$\gamma_t^u(h) = \phi^{u_1}\left(e^{\phi^{u_2}(u_t(h)-1)} - 1\right), \tag{8}$$

where ϕ_1^u and ϕ_2^u are parameters determining the cost of deviating from the steady state utilization rate (normalized to one). Firm h's law of motion for physical capital reads:

$$K_t(h) = (1-\delta) K_{t-1}(h) + \kappa_t(h) K_{t-1}(h), \tag{9}$$

where $\delta \in [0,1]$ is the rate of depreciation and $\kappa_t(h)$ denotes capital adjustment costs. The latter takes the following form:

$$\begin{aligned}\kappa_t(h) \;=\;& \frac{I_t(h)}{K_{t-1}(h)} - \frac{\phi_1^I}{2}\left[\left(\frac{I_t(h)}{K_{t-1}(h)} - \frac{I}{K}\right)\right]^2 \\ & -\frac{\phi_2^I}{2}\left(\frac{I_t(h)}{K_{t-1}(h)} - \frac{I_{t-1}}{K_{t-2}}\right)^2 + z_t^I,\end{aligned} \tag{10}$$

where I_t denotes investment and z_t^I is an AR(1) investment shock with autoregressive parameter λ^I and standard deviation σ_I. The labor input is a CES aggregate of hours supplied by the different households:

$$l_t(h) = \left[\int_0^1 l_t(h,j)^{1-\frac{1}{\psi_t}} dj\right]^{\frac{\psi_t}{\psi_t - 1}}, \tag{11}$$

where ψ_t is an AR(1) process governing the elasticity of substitution between different types of labor with autoregressive parameter λ^ψ and standard deviation σ_ψ. Firms sell their goods in markets characterised by monopolistic competition. International goods markets are segmented and firms set prices in the local currency of the buyer. An individual firm h charges $P_t^Q(h)$ in the home market and $P_t^{M^*}(h)$ abroad, where the latter is denoted in foreign currency. The quantity sold in domestic markets is denoted $Q_t(h)$ and exports are denoted $M_t^*(h)$. Nominal price stickiness is modelled by assuming that firms face quadratic costs of adjusting prices,

$$\gamma_t^{PQ}(h) \equiv \frac{\phi^Q}{2}\left[\frac{P_t^Q(h)}{\pi P_{t-1}^Q(h)}-1\right]^2+\frac{\phi_2^Q}{2}\left[\frac{P_t^Q(h)/P_{t-1}^Q(h)}{P_{t-1}^Q/P_{t-2}^Q}-1\right]^2 \text{ and} \tag{12}$$

$$\gamma_t^{PM^*}(h) \equiv \frac{\phi^{M^*}}{2}\left[\frac{P_t^{M^*}(h)}{\pi P_{t-1}^{M^*}(h)}-1\right]+\frac{\phi^{M^*}}{2}\left[\frac{P_t^{M^*}(h)/P_{t-1}^{M^*}(h)}{P_{t-1}^{M^*}/P_{t-2}^{M^*}}-1\right]^2, \tag{13}$$

in the domestic and foreign market, respectively, where π is the steady-state inflation rate. Firms choose hours, capital, investment, the utilization rate and prices to maximize present discounted value of cash-flows, adjusted for the intangible cost of changing prices, taking into account the law of motion for capital, and demand both at home and abroad. The foreign intermediate goods sector is modelled symmetrically. The output gap and marginal costs in the foreign economy are modelled as AR(1) processes with autoregressive parameters λ^{y^*} and λ^{mc^*} and standard deviations σ_{y^*} and σ_{mc^*},respectively.

Households The economy is inhabited by a continuum of infinitely-lived households indexed by $j \in [0,1]$. The lifetime expected utility of household j is:

$$U_t(j) = E_t \sum_{i=0}^{\infty} \beta^i \left[u\left(C_{t+i}(j)\right) - v\left(l_{t+i}(j)\right)\right], \tag{14}$$

where C denotes consumption, l is hours worked and β is the discount factor $0 < \beta < 1$. The current period utility functions, $u(C_t(j))$ and $v(l_t(j))$, are

$$u\left(C_t(j)\right) = \left(1 - b^c/\pi^z\right) \ln\left[\frac{\left(C_t(j) - b^c C_{t-1}\right)}{1 - b^c/\pi^z}\right], \tag{15}$$

and

$$v\left(l_t(j)\right) = \frac{1}{1+\zeta} l_t(j)^{1+\zeta}. \tag{16}$$

where $\zeta > 0, b^c$ $(0 < b^c < 1)$ governs the degree of habit persistence and π_z denotes the steady-state growth rate in the economy. Each household is the monopolistic supplier of a differentiated labor input and sets the nominal wage subject to the labor demand of intermediate goods firms and subject to quadratic costs of adjustment, γ^W:

$$\gamma_t^W(j) \equiv \frac{\phi^W}{2}\left[\frac{W_t(j)}{\pi_w W_{t-1}(j)} - 1\right]^2 + \frac{\phi_2^W}{2}\left[\frac{W_t(j)/W_{t-1}(j)}{W_{t-1}/W_{t-2}} - 1\right]^2 \tag{17}$$

where W_t is the nominal wage rate and π_w is the steady-state growth rate of nominal wages. The individual flow budget constraint for agent j is:

$$\begin{aligned} P_t C_t(j) + S_t B_{H,t}^*(j) + B_t(j) \leq W_t(j)\, l_t(j)\left[1 - \gamma_t^W(j)\right] \\ + \left[1 - \gamma_{t-1}^{B^*}\right]\left(1 + r_{t-1}^*\right) S_t B_{H,t-1}^*(j) \\ + \left(1 + r_{t-1}\right) B_{t-1}(j) + DIV_t(j) - TAX_t(j), \end{aligned} \tag{18}$$

where S_t is the nominal exchange rate, $B_t(j)$ and $B_{H,t}^*(j)$ are household j's end of period t holdings of domestic and foreign bonds, respectively. Only the latter are traded internationally. The domestic short-term nominal interest rate is denoted by r_t, and the nominal return on foreign bonds is r_t^*. The variable DIV includes all profits from intermediate goods firms and nominal wage adjustment costs, which are rebated in a lump-sum fashion. Home agents pay lump-sum net taxes, TAX_t, denominated in home currency. The financial intermediation cost takes the following form:

$$\gamma_t^{B^*} = \phi^{B1}\frac{\exp\left(\phi^{B2}\left(\frac{S_t B_t^{H^*}}{P_t Z_t}\right)\right) - 1}{\exp\left(\phi^{B2}\left(\frac{S_t B_t^{H^*}}{P_t Z_t}\right)\right) + 1} + z_t^B, \tag{19}$$

where $0 \leq \phi^{B1} \leq 1$ and $\phi^{B2} > 0$ and where the 'risk premium', z_t^B, is assumed to follow an AR(1) process with autoregressive parameter λ^B and standard deviation σ_B.

Government The government purchases final goods financed through a lump-sum tax. Real government spending (adjusted for productivity), $g_t \equiv G_t/Z_t$, is modelled as an AR(1) process with autoregressive parameter λ^g and standard deviation σ_g. The central bank sets the interest rate according to a simple instrument rule, which in its log-linearised version takes the form:

$$r_t = \omega_r r_{t-1} + (1-\omega_r)\left[\begin{array}{c} \omega_\pi \pi_{t-1} + \omega_y \widehat{y}_{t-1} + \omega_{rer} rer_{t-1} \\ +\omega_{\Delta\pi}\left(\pi_{t-1} - \pi_{t-2}\right) + \omega_{\Delta y} \Delta\widehat{y}_{t-1} \end{array}\right] + z_t^r \tag{20}$$

where π_t is the aggregate inflation rate, rer_t is the (log) real exchange rate and z_t^r is a mean-zero monetary policy shock with standard deviation σ_r. The parameter $\omega_r \in [0, 1\rangle$ determines the degree of interest rate smoothing. The output gap $\widehat{y}_t$ is measured as the percentage deviation of gross domestic product Y_t from the stochastic productivity trend. The remaining variables are in deviation from their steady-state levels.

Table 3: Variable definitions and sources

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
π_t	Overall inflation adjusted for taxes and excl. energy prices (CPI-ATE), s.a.	Statistics Norway
π_t^m	Imported inflation 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

Table 4: Calibrated parameter values and prior distributions

Calibrated parameters	Value	Estimated parameters	Prior		Estimated parameters	Prior	
			Type	Mean (std)		Type	Mean (std)
θ	6.0000	α	Beta	0.3000 (0.020)	σ_z	Inv gam	0.0050 (Inf)
θ^*	6.0000	ψ	Inv gam	5.5000(0.500)	σ_r	Inv gam	0.0025 (Inf)
π^z	1.0056	ζ	Inv gam	3.0000 (0.200)	σ_ψ	Inv gam	1.0000 (Inf)
β	0.9993	μ	Inv gam	1.1000 (0.200)	σ_θ	Inv gam	1.0000 (Inf)
ϕ^{I1}	1.0000	μ^*	Inv gam	1.1000 (0.200)	σ_I	Inv gam	1.0000 (Inf)
ϕ^{B1}	1.0000	b_C	Beta	0.7500 (0.050)	σ_L	Inv gam	0.0050 (Inf)
δ	0.0180	ϕ^M	Inv gam	1.0000 (1.000)	σ_B	Inv gam	0.0100 (Inf)
η	0.6444	ϕ_2^M	Inv gam	1.0000 (1.000)	σ_{mc^*}	Inv gam	0.0100 (Inf)
ξ	0.7000	ϕ^Q	Inv gam	1.0000 (1.000)	σ_{y^*}	Inv gam	0.0100 (Inf)
ϕ_1^U	1.0000	ϕ_2^Q	Inv gam	1.0000(1.000)	σ_g	Inv gam	0.0100 (Inf)
ϕ^{M^*}	0.5000	ϕ^W	Inv gam	1.0000 (1.000)	λ^z	Beta	0.8500 (0.1)
$\phi_2^{M^*}$	0.0000	ϕ_2^W	Inv gam	1.0000 (1.000)	λ^ψ	Beta	0.8500 (0.1)
		ϕ_2^I	Gam	10.0000 (5.000)	λ^θ	Beta	0.8500 (0.1)
		ϕ_2^B	Inv gam	0.0200 (0.005)	λ^I	Beta	0.8500 (0.1)
		ϕ_2^U	Inv.gam	0.5000 (0.020)	λ^L	Beta	0.8500 (0.1)
		ω_r	Beta	0.8000 (0.050)	λ^B	Beta	0.8500 (0.1)
		ω_π	Inv.gam	2.0000 (0.100)	λ^{mc^*}	Beta	0.8500 (0.1)
		ω_y	Normal	0.2000 (0.050)	λ^{y^*}	Beta	0.8500 (0.1)
		ω_{rer}	Normal	0.0000 (0.050)	λ^g	Beta	0.8500 (0.1)
		$\omega_{\Delta y}$	Normal	0.0000 (0.050)			
		$\omega_{\Delta\pi}$	Normal	0.0000 (0.050)			