



Conference on

Forecasting and Monetary Policy

Berlin, 23-24 March 2009

Anne Sofie Jore
Central Bank of Norway

James Mitchell
NIESR London

Shaun Vahey
Melbourne Business School, Central Bank of Norway and Reserve
Bank of New Zealand

Ida Wolden Bache
Central Bank of Norway

**„Combining Forecast Densities from VARs and
DSGEs with Uncertain Instabilities“**

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)

February 27, 2009

Abstract

A popular macroeconomic forecasting strategy takes combinations across many models to hedge against model instabilities of unknown timing; see (among others) Stock & Watson (2004), Clark & McCracken (2009), and Jore, Mitchell and Vahey (2009). The scope of such ensemble forecasting exercises usually excludes Dynamic Stochastic General Equilibrium (DSGE) models, such as those advocated by Del Negro and Schorfheide (2004) and Smets and Wouters (2007), limiting the computational burden. In this paper, we use an expert combination framework (Winkler, 1981) to combine forecast densities from Vector Autoregressions (VARs), and a DSGE model (NEMO: the Norges Bank core policymaking macromodel). We show that the predictive densities from the DSGE model are competitive with those from a VAR ensemble if the VAR components are restricted to have constant parameters. In this case, both the VAR ensemble and the DSGE forecast densities are poorly calibrated. However, a VAR ensemble which encompasses structural break components produces well-calibrated forecast densities. The VARs with breaks are relatively responsive to structural breaks of unknown timing.

Keywords: Forecast densities; Ensemble forecasting; Evaluating forecasts; VAR models; DSGE models

JEL codes: C32; C53; E37

*Contact: James Mitchell, NIESR, 2 Dean Trench Street, Smith Square, London, SW1P 3HE, U.K. Tel: +44 (0) 207 654 1926. Fax: +44 (0) 207 654 1900. E-Mail: j.mitchell@niesr.ac.uk. 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 Heather Anderson, Fabio Canova, Hugo Gerard, Adrian Pagan and the participants at the Monetary Policy in Open Economies Workshop, Reserve Bank of Australia, December 2007.

1 Introduction

An increasingly common macroeconomic forecasting strategy takes combinations across forecasts from many models to hedge against model instabilities of unknown timing. Recent studies include (among others) Stock & Watson (2004), Clark & McCracken (2009), and Jore et al. (2009). The first two studies 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 densities.

The existing ensemble macro forecasting literature excludes Dynamic Stochastic General Equilibrium (DSGE) models (as seen, for example, in Del Negro & Schorfheide (2004) and Smets & Wouters (2007)), in part because of the computational burden imposed by both ensemble and DSGE forecasting. The absence of DSGE models presents a practical difficulty for the implementation of ensemble forecasting methods at central banks. DSGE models are the dominant tools for monetary policymakers conducting structural analysis.

In this paper, we use an expert combination framework (Winkler (1981)) to combine forecast densities from Vector Autoregressions (VARs) and a DSGE model (NEMO, the Norges Bank core policymaking model). Component forecasts are combined to produce the ensemble forecast density using the logarithmic score; see (among others) Jore et al. (2009). We evaluate the forecast densities using probability integral transforms (*pits*). This offers a means of evaluating density forecasts for general but unknown loss functions. We find the predictive densities from the DSGE model and those from an ensemble of VARs with constant parameters to be poorly calibrated. However, expanding the model space for the VAR ensemble to encompass structural break components gives well-calibrated densities.

We draw the following conclusions. First, ensemble densities based on component VARs with breaks are reliable, whereas ensembles from constant parameter VARs are not. Second, although the policy DSGE model studied here produces accurate point forecasts (relative to a benchmark constant parameter VAR), the forecast densities are poorly calibrated. Our interpretation is that central banks using DSGE models to produce forecast densities (also known as “fan charts”) should consider forecast combination as a way of producing well-calibrated predictive densities. Or, they should develop DSGE variants that adapt better to structural breaks.

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, in Section 4 we discuss the Norwegian sample. 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 for the combinations of a large number of models using forecast density combination methods. Earlier papers, by Jore et al. (2009) and Garratt et al. (2009), take this approach to ensemble modeling in examining predictive densities from VAR models using US data. In contrast, Stock & Watson (2004) and Clark & McCracken (2009) study point forecast combinations across a large number of models. Although point forecast combination has a longer tradition in economics (e.g., see Bates & 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. We note that many central banks, including Norges Bank, provide forecast densities to communicate the policy stance.

2.1 Forecast density combination

The ensemble densities are defined by the convex combination:¹

$$p(Y_{\tau,h}) = \sum_{i=1}^N w_{i,\tau,h} g(Y_{\tau,h} | I_{i,\tau}), \quad \tau = \underline{\tau}, \dots, \bar{\tau}, \quad (1)$$

where $g(Y_{\tau,h} | I_{i,\tau})$ are the h -step ahead forecast densities from component model i , $i = 1, \dots, N$ of a random variable Y_{τ} , (with realisation y_{τ}), conditional on the information 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 $\tau = \underline{\tau}, \dots, \bar{\tau}$. 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 & Giacomini (2007) and Hall & Mitchell (2007), we use the logarithmic score to measure density fit for each component model through the evaluation period. The logarithmic scoring rule gives a high score to a density forecast that assigns a high probability to the realised value.³ Specifically, following Jore et al. (2009) the

¹The linear opinion pool is sometimes justified by considering an expert combination problem. See for example, Morris (1974, 1977) and Winkler (1981), Lindley (1983) and McConway (1990). Wallis (2005) proposes the linear opinion pool as a tool to aggregate forecast densities from survey participants. Mitchell & Hall (2005) combine inflation density forecasts from two institutions.

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

³The logarithmic score of the i -th density forecast, $\ln g(y_{\tau,h} | I_{i,\tau})$, is the logarithm of the probability density function $g(\cdot | I_{i,\tau})$, evaluated at the outturn $y_{\tau,h}$.

recursive weights for the h -step ahead densities take the form:

$$w_{i,\tau,h} = \frac{\exp \left[\sum_{\underline{\tau}-10}^{\tau-1-h} \ln g(y_{\tau,h} | I_{i,\tau}) \right]}{\sum_{i=1}^N \exp \left[\sum_{\underline{\tau}-10}^{\tau-1-h} \ln g(y_{\tau,h} | I_{i,\tau}) \right]}, \quad \tau = \underline{\tau}, \dots, \bar{\tau} \quad (2)$$

where the $\underline{\tau}-10$ 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 .

From a Bayesian perspective, density combination based on recursive logarithmic score weights, RLSW, has many similarities with an approximate predictive likelihood approach (see Raftery & Zheng (2003), and Eklund & Karlsson (2007)).⁴ 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. Andersson & Karlsson (2007) propose (informative) Bayesian predictive likelihood methods for VAR forecast combination but do not consider forecast density evaluation.⁵

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 using the probability integral transforms (*pits*) of the realisation 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 conditionally calibrated; i.e., if the *pits* $z_{\tau,h}$, where $z_{\tau,h} = \int_{-\infty}^{y_{\tau,h}} p(u) du$, 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.⁶

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 $h = 1$, 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,

⁴These similarities are lost for $h > 1$.

⁵Beyond the economics literature, Raftery et al. (2005) and Carvalho & Tanner (2006) exploit the EM algorithm to estimate weights and the parameters of their component densities. Given the computational burden imposed by the DSGE model under consideration in the current application, we leave this avenue for future research.

⁶Given the large number of VAR component densities under consideration, we do not allow for parameter uncertainty when evaluating the *pits*. Corradi & Swanson (2006) review *pits* tests computationally feasible for small N .

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.⁷ For $h > 1$ we test for autocorrelation at lags greater than $(h - 1)$, but less than $n = 6$, using a modified LB test (MLB).⁸ Even for correctly calibrated densities, we expect autocorrelation stemming from the overlapping forecast horizons.

Since more than one density forecast can produce uniform and even independent *pits*, in order to discriminate between them, as suggested by Mitchell & Wallis (2009), we also consider Kullback-Leibler information criterion (KLIC)-based tests. They involve using a LR test, based on the difference between the logarithmic score of each forecast and the forecast with the highest score, to test equal predictive accuracy based on the two densities' KLIC difference. Amisano & Giacomini (2007) develop the same test, starting with the logarithmic score as a measure of forecast performance.

3 Component models

3.1 VARs

We consider a range of VAR models in inflation, output growth and the interest rate, with lag lengths of 1 to 4, and selected by the BIC. For each specification, we estimate trivariate VARs, bivariate VARs and ARs (always including inflation). For the (trivariate) VARs we also transform the variables prior to estimation in two ways: we include first-differenced VARs (DVARs), and de-trended VARs (using an exponential smoother). This gives 30 models in total.

We utilise a direct forecast methodology (see Marcellino et al. (2003)) to generate the h -step ahead predictive densities from each VAR (AR). Consider a single equation from

⁷To 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.

⁸With $T = (\bar{\tau} - \underline{\tau})$, the number of observations in the evaluation period

$$MLB = \frac{(T + 2) \sum_{j=h}^n (T - j)^{-1} \hat{\rho}_j^2}{\left(1 + 2 \sum_{j=1}^{h-1} \hat{\rho}_j^2\right) / T}$$

where $\hat{\rho}_j$ is the sample autocorrelation at lag j and $MLB \sim \chi_{n-(h-1)}^2$ under the null hypothesis of no serial correlation between lags h and n .

a given (constant parameter) VAR model

$$Y_t = \alpha + \beta X_{t-h} + \sigma \varepsilon_t, \quad (3)$$

where $t = 1, \dots, \tau$ refers to the sample used to fit the model, $\varepsilon_t \sim i.i.d. N(0, 1)$ and the parameters are collected in α , β and σ . The predictive densities for $Y_{\tau+h}$ (with non-informative priors), allowing for small sample issues, are multivariate Student-t; see Zellner (1971), pp. 233-236 and, for a more recent application, Garratt, Koop, Mise & Vahey (2008).

Following Jore et al. (2009) and Garratt, Koop & Vahey (2008) we deal with structural breaks, in the conditional mean and/or variance, by estimating a given VAR model, for a given recursion, with each candidate break date. Thereby we accommodate break-date uncertainty in a convenient manner. For computational simplicity, we restrict the break dates to be identical across equations for each VAR model, and consider every feasible break date value with a regime containing at least 40% of the observations. Even so, with new break models included for every four recursions in the evaluation period (thereby accommodating breaks each year rather than each quarter), the computational burden is considerable. Hence, we further restrict the maximum number of regimes, R , to 3. With these additional structural break models added to the set of full sample VARs, we consider a maximum of 810 component models (for the final recursion in our evaluation period). We also estimate the 30 models over a rolling window of 34 quarters (24 quarters for the ARs) to give 840 component models in total. Rolling regression models are advocated as a means of tackling structural breaks by (among others) Eklund et al. (2008).

A VAR ensemble is constructed by combining the component densities, weighted using RLSW and a 10 observation “training period” prior to the evaluation. We construct VAR ensembles based on VAR components both with constant parameters and allowing for breaks. We refer to these as a “no break VAR ensemble” and a “break VAR ensemble”, respectively.

3.2 The DSGE model: NEMO

NEMO (the Norwegian Economy MOdel) is the core model used for structural analysis at Norges Bank. It is a medium-scale New Keynesian small open economy model with a similar structure to the DSGE models recently developed in many central banks, e.g., Sveriges Riksbank (see Adolfson et al. (2008)).⁹ 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 mod-

⁹See Brubakk et al. (2006) for a more thorough discussion of NEMO and the DSGE literature.

ifications to the Bayesian simulation methodology and the steady-state behaviour of the model.

An appendix describes the NEMO economy in detail. Here we summarise 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 utilisation 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.

Estimation uses data on the following eleven 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, the overnight deposit rate (the policy rate) and hours worked.¹⁰ 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. The sample used for estimation starts in 1987Q1 to match the practice used in Norges Bank monetary policymaking applications.

We estimate the structural parameters using Bayesian techniques.¹¹ The forecast draws are based on the mode of the posterior distributions for the structural parameters; the forecast densities (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.

¹⁰The national accounts data relate to the mainland economy, that is, the total economy excluding the petroleum sector.

¹¹We carry out DSGE estimation in DYNARE; Karagedikli et al. (2007) provide a simplified DSGE example with code.

4 The Norwegian data

In examining the predictives from the DSGE model and the VARs (and their combinations), we restrict our combinations and evaluations to inflation. We measure inflation as the headline consumer price index adjusted for tax changes and excluding temporary changes in energy prices—the underlying rate. Recall that the DSGE model uses 11 observable for estimation and that the VARs include up to three variables (where those variables are: the 3-month money market rate and GDP, excluding oil and gas sectors, seasonally adjusted.) Hence we are examining the predictive performance of the components and ensembles using limited information. Our decision to focus on inflation is partly motivated by the inflation targeting regime adopted by many central banks, including Norges Bank.¹²

Our recursive forecasting exercise is intended to mimic the behaviour 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 utilise real-time macroeconomic data because the data have not been compiled for (most of) the 11 observables used in DSGE estimation. Instead, we use a single vintage of data available in 2008Q4 for all forecasts and realisations.¹³

The recursive forecast experiments are constructed as follows: We estimate each model on a sample ending in 1997Q2 and compute forecasts for the endogenous variables (up to 3 variables for VARs, 11 for the DSGE) for horizons of one up to four quarters. We construct (recursive) ensemble predictive densities for this observation in the manner described in Section 2. Then we extend the sample by one quarter, re-estimate the models, compute new forecasts for each component model, and produce the ensemble predictives. This exercise is repeated until the end of the sample, 2007Q3. The evaluation period starts in 1997Q3 for one step ahead forecasts, in 1997Q4 for two step ahead forecasts and so on. For all horizons, the evaluation period ends in 2007Q3.

5 Results

Prior to evaluating the forecast densities using the *pits* tests, we consider the RMSFE of the VAR and DSGE ensembles. Table 1 lists the RMFSE statistics for both the no break and break VAR variants, as well as the DSGE model. There is little to choose between the constant parameter and break VAR ensembles at shorter horizons. At longer horizons (e.g., $h = 4$), the evidence suggests that the break VAR ensemble is forecasting

¹²The forecast densities for output growth also appear well-calibrated at a 95% significance level for the break VAR ensemble (available from the authors upon request). The no break VAR ensemble and the DSGE predictives are not.

¹³Current Norges Bank research aims to build an appropriate real-time database.

more accurately. However, at all horizons, the DSGE predictive densities give a higher RMFSE than both the break and no break VAR ensembles. However, we note that these differences are not statistically significant, at a 95% level, using Diebold-Mariano tests.

A couple of remarks on these findings. First, as other studies have found for VAR models, the ensemble approach is effective at producing accurate point forecasts. Second, the DSGE model is not as effective as the ensemble of VARs, with or without breaks. We emphasise that many studies have found that DSGE models produce out of sample point forecasts competitive with (constant parameter) autoregressive models; see, for example, Adolfson et al. (2008) and Schorfheide et al. (2008). However, these earlier studies did not examine the forecast performance of VAR ensembles. In common with the findings elsewhere in the literature, the out-of-sample forecasting accuracy of the DSGE model is competitive with a single constant parameter VAR representation.

Given our focus on performance of the ensemble, rather than its components, we do not report the individual model weights (RLSW) for the many structural break variants in the break VAR ensemble. But we do note that the break models with the most support typically have a single break in the interval 1985 to 1987. As Figure 1, which plots the annualised quarterly inflation rate, shows over this interval there was a marked reduction in both the level and the volatility of inflation in Norway, with the government focusing on stabilising inflation via a stable exchange-rate, higher interest-rates, fiscal restraint and wage “understandings”.

Table 1: RMSFE statistics

Horizon	AR(1)	DSGE	break VAR ensemble	no break VAR ensemble
$h = 1$	0.80	0.86	0.77	0.77
$h = 2$	0.81	1.01	0.82	0.83
$h = 3$	0.95	1.07	0.84	0.88
$h = 4$	1.01	1.23	0.94	1.02

In Table 2 we turn to the *pits* tests on the forecast densities. To facilitate easy-reading, we place the p -values, or test statistics, in bold when the density forecast is correctly calibrated at a 95% significance level—that is, when we cannot reject the null hypothesis that the densities are correctly calibrated according to one of the four evaluation tests.¹⁴ We also report, in the final column of Table 2, the average logarithmic score ($\log S$) of each density forecast; the logarithmic score serves as the basis for the KLIC-based tests used to discriminate between competing forecasts.

Table 2 shows that at shorter horizons ($h = 1$ and $h = 2$), the densities from the no

¹⁴To 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 95% significance level, of $(100\% - 95\%)/4 = 1.25\%$ rather than 5% on each test. This gives qualitatively similar findings.

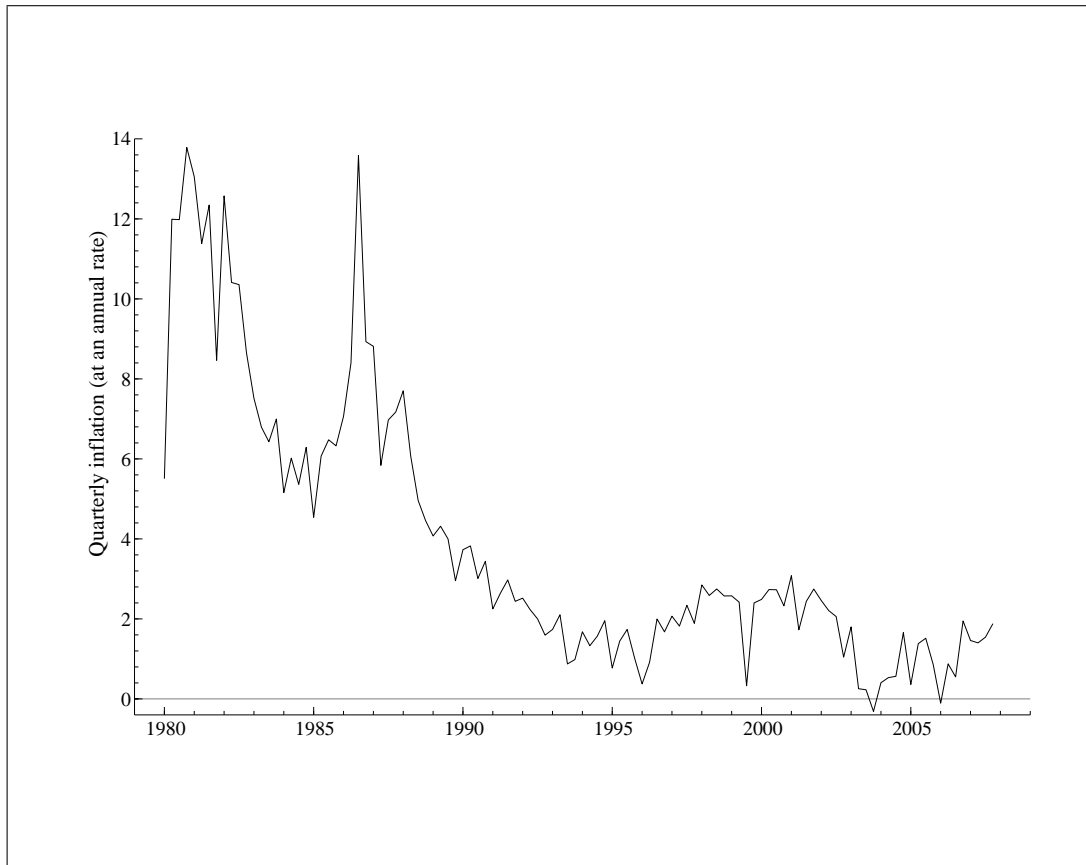


Figure 1: Norwegian inflation

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 no calibration error across the four tests.

The DSGE densities perform much like the no break VAR ensemble at shorter horizons—that is, they are poorly calibrated according to the *pits* tests. Unlike the constant parameter ensemble, calibration is also poor at longer horizons. The break VAR ensemble densities, since they give substantial weight to models that allow for the structural breaks in the mid 1980s, are well calibrated at 95% at all 4 horizons. The break VAR ensemble also has the highest average logarithmic score.

The preceding analysis has focused on evaluating the calibration of the ensemble VAR and DSGE predictive densities. It is also instructive to combine these two predictives using the recursive log scores. Garratt et al. (2009) refer to a combination of ensemble predictive densities as a “grand ensemble”. In our application, one of the candidates is not an ensemble—the DSGE predictive density—but the methodology is similar.¹⁵

Figure 2 plots the time variation in the weights on the two grand ensembles; for the

¹⁵Our experiment could be re-constructed with the DSGE model evaluated as if it *were* a candidate VAR ensemble component. With at most 3 variables in the VAR and 11 in the DSGE, the DSGE and the VARs do not have a similar model space.

no break VARs and the DSGE, and break VARs and the DSGE, respectively. Inspection of these weights indicates that the DSGE densities are competitive, at shorter horizons ($h = 1$ and $h = 2$), against the no break VAR ensemble. While the weight on the DSGE varies over time, at both $h = 1$ and $h = 2$ the DSGE density receives on average (across the evaluation period) a weight of roughly 0.5 at $h = 1$ and 0.3 at $h = 2$. However, the weight on the DSGE density in the grand ensemble with the break VARs is much lower. In this case, the DSGE density receives on average a weight of 0.1 (at $h = 1$) and 0.0 (at $h = 2$). Moreover, by the end of the evaluation period, the RLSW weight on the DSGE density is approximately zero at both $h = 1$ and $h = 2$.

At longer horizons the DSGE density receives a smaller weight than at shorter horizons. The RLSW on the DSGE reaches zero half way through the evaluation period at both $h = 3$ and $h = 4$. This happens even in combination with the no break VAR. This is consistent with the *pits* evidence from Table 2, which shows at longer horizons ($h = 3$ and $h = 4$) the densities from the no break VARs to be well calibrated.

Finally, Table 3 contains the *pits* tests for the grand ensemble predictive densities. We note that both cases, based on combining the DSGE with the break VARs, or with the no break VARs, display good calibration. But, at shorter horizons, the grand ensemble based on the no break VARs and the DSGE (grand ensemble I) displays calibration failure for some tests. Using the KLIC-based tests of equal predictive accuracy, we cannot reject the null hypothesis that the grand ensemble based on the break VARs and the DSGE (grand ensemble II) performs equally as well as the break VAR ensemble considered in Table 2.

We emphasise that throughout our analysis we are using the component recursive logarithmic score, RLSW, to construct the VAR ensemble predictives. An alternative approach described by Hall & Mitchell (2007) and Geweke & Amisano (2008) is to find 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. But we checked our findings for the grand ensemble (where there are only two predictive densities to be combined in each case), and found the weights to be similar to those reported above. The exceptions are at $h = 3$ and $h = 4$ where, from the beginning of the evaluation period rather than half way through it as in Figure 2 for the RLSW, the iterative weight on the DSGE density is zero.

6 Conclusions

We draw the following conclusions from our evaluations of the forecast densities. First, ensemble densities based on component VARs with breaks are well calibrated. Second, ensembles from constant parameter VARs are typically poorly calibrated. Third, although many policymakers prefer DSGE models for structural analysis, the forecast densities do

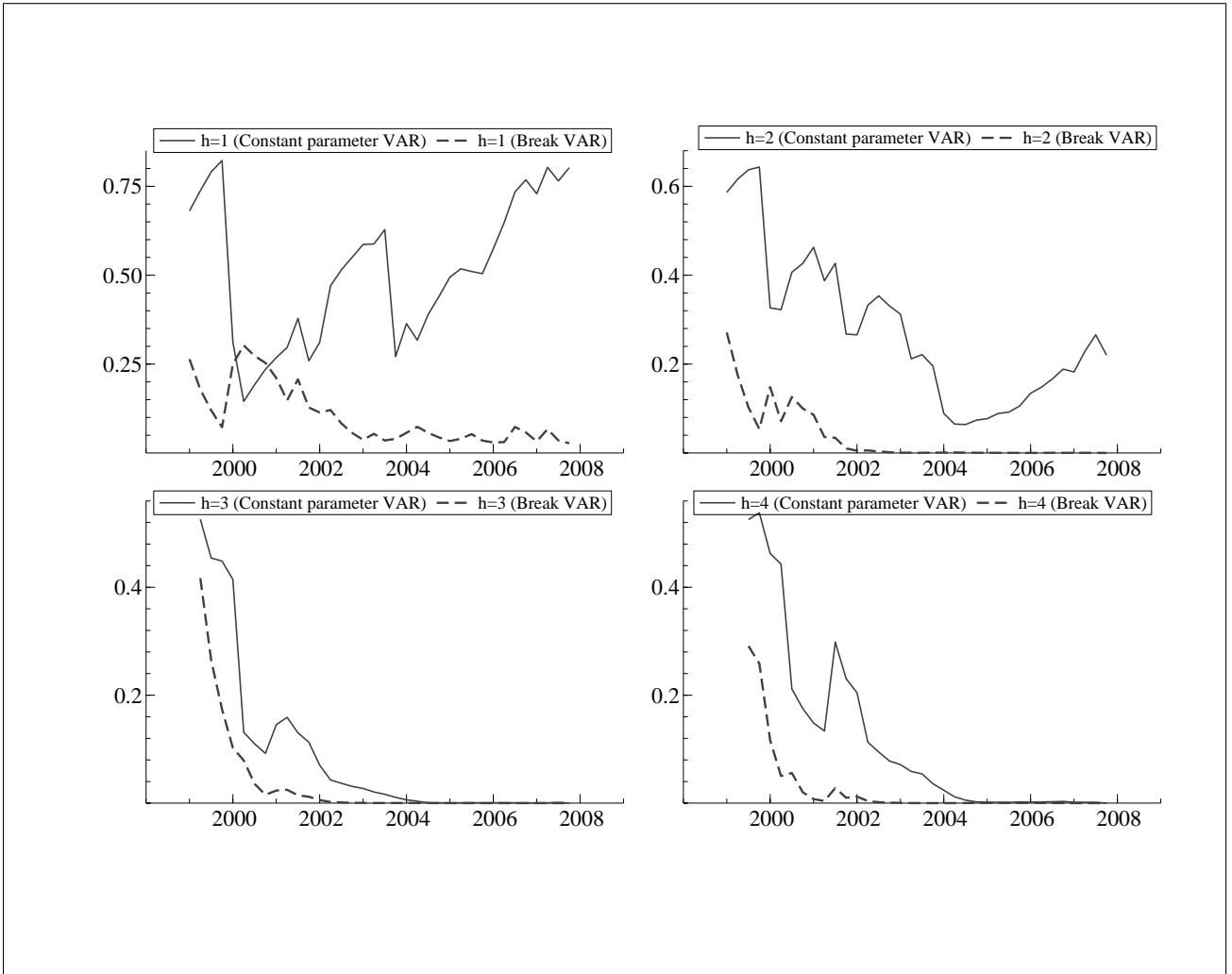


Figure 2: Recursive logarithmic score weights on the DSGE in the grand ensembles

not match the performance of break VAR ensembles. And the DSGE model receives a low weight in the grand ensemble with break VARs. When combined with constant parameter VARs, the densities from the DSGE receive a substantial weight at some horizons.

Clearly central banks use DSGE models because they have appealing theoretical properties, appropriate for policy analysis. In this study, we have restricted our attention to forecasting properties. Our finding that the forecast densities produced by DSGEs are poorly calibrated will spur further development in DSGE modeling. In particular, it seems that DSGE models which allow for time variation in the parameters offer the scope for better predictive densities. Some recent candidates include Del Negro et al. (2007), Fernandez-Villaverde & Rubio-Ramirez (2007), and Justiniano & Primiceri (2008). To our knowledge, no central banks have yet adapted this type of DSGE model for policy use. In their absence, ensemble forecast combination offers a reliable methodology for producing well-calibrated forecast densities.

Table 2: Density forecast evaluation using the *pits*

$h = 1$	LR3	AD	χ^2	LB1	logS
no break VAR ensemble	0.00	1.80	0.01	0.56	-1.287
break VAR ensemble	0.43	0.96	0.13	0.83	-1.185
DSGE	0.00	3.27	0.02	0.02	-1.265
$h = 2$	LR2	AD	χ^2	MLB	logS
no break VAR ensemble	0.00	2.61	0.01	0.86	-1.390
break VAR ensemble	0.37	1.16	0.32	0.94	-1.223
DSGE	0.00	7.11	0.00	0.82	-1.440*
$h = 3$	LR2	AD	χ^2	MLB	logS
no break VAR ensemble	0.08	0.83	0.20	0.75	-1.388
break VAR ensemble	0.36	1.07	0.49	0.85	-1.319
DSGE	0.00	9.36	0.00	0.91	-1.590*
$h = 4$	LR2	AD	χ^2	MLB	logS
no break VAR	0.10	0.77	0.87	0.68	-1.537
break VAR	0.06	2.36	0.12	1.00	-1.487
DSGE	0.00	9.15	0.00	0.93	-1.706

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 Anderson-Darling test statistic for uniformity of the *pits* which assuming independence of the *pits* has an associated 95 percent asymptotic critical value of 2.5. 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 ; logS is the average logarithmic score and an asterisk denotes rejection at 95%, using the KLIC-based LR test, of equal predictive accuracy of the ensemble concerned against the ensemble with the highest logarithmic score

Table 3: Density forecast evaluation of the grand ensembles using the *pits*

$h = 1$	LR3	AD	χ^2	LB1	logS
grand ensemble I	0.01	2.08	0.42	0.48	-1.270
grand ensemble II	0.33	1.16	0.06	0.65	-1.193
$h = 2$	LR2	AD	χ^2	MLB	logS
grand ensemble I	0.00	3.43	0.01	0.91	-1.411*
grand ensemble II	0.32	1.32	0.40	0.91	-1.226
$h = 3$	LR2	AD	χ^2	MLB	logS
grand ensemble I	0.08	0.94	0.08	0.78	-1.413
grand ensemble II	0.31	1.19	0.60	0.81	-1.332
$h = 4$	LR2	AD	χ^2	MLB	logS
grand ensemble I	0.05	1.21	0.82	0.70	-1.561
grand ensemble II	0.07	2.39	0.12	1.00	-1.487

Notes: grand ensemble I refers to the combination of no break VARs and the DSGE; grand ensemble II refers to break VARs and the DSGE. Also see notes to Table 2. The KLIC-based LR tests for equal predictive accuracy are performed relative to the forecast with the highest logarithmic score in Table 2

References

- Adolfson, M., Lasen, S., Lind, J. & Villani, M. (2008), ‘Evaluating an estimated new keynesian small open economy model’, *Journal of Economic Dynamics and Control* **32**(8), 2690 – 2721.
- Amisano, G. & Giacomini, R. (2007), ‘Comparing density forecasts via weighted likelihood ratio tests’, *Journal of Business and Economic Statistics* **25**, 177–190.
- Andersson, M. & Karlsson, S. (2007), Bayesian forecast combination for VAR models. Unpublished manuscript, Sveriges Riksbank.
- Bates, J. M. & Granger, C. W. J. (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., Husebø, T. A., Maih, J., Olsen, K. & Østnor, M. (2006), Finding NEMO: Documentation of the norwegian economy model, Norges Bank Staff Memo 2006/6.
- Carvalho, A. X. & Tanner, M. A. (2006), ‘Modeling nonlinearities with mixtures-of-experts of time series’, *International Journal of Mathematics and Mathematical Sciences* **2006**(9), 1–22.
- Clark, T. E. & McCracken, M. W. (2009), ‘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.
- Corradi, V. & Swanson, N. R. (2006), Predictive density evaluation, in G. Elliott, C. W. J. Granger & A. Timmermann, eds, ‘Handbook of Economic Forecasting’, North-Holland, North Holland, pp. 197–284.
- Dawid, A. P. (1984), ‘Statistical theory: the prequential approach’, *Journal of the Royal Statistical Society B* **147**, 278–290.
- Del Negro, M. & Schorfheide, F. (2004), ‘Priors from general equilibrium models for VARs’, *International Economic Review* **45**, 643–673.
- Del Negro, M., Schorfheide, F., Smets, F. & Wouters, R. (2007), ‘On the fit of new Keynesian models’, *Journal of Business & Economic Statistics* **25**, 123–143.

- Diebold, F. X., Gunther, A. & Tay, K. (1998), ‘Evaluating density forecasts with application to financial risk management’, *International Economic Review* **39**, 863–883.
- Eklund, J., Kapetanios, G. & Price, S. (2008), Forecasting in the presence of recent structural breaks. Presented at the Nowcasting with Model Combination Workshop, RBNZ, December 2008.
- Eklund, J. & Karlsson, S. (2007), ‘Forecast combination and model averaging using predictive measures’, *Econometric Reviews* **26**(2-4), 329–363.
- Fernandez-Villaverde, J. & Rubio-Ramirez, J. (2007), ‘Estimating macroeconomic models: A likelihood approach’, *Review of Economic Studies* **74**(4), 1059–1087.
- Garratt, A., Koop, G., Mise, E. & Vahey, S. P. (2008), ‘Real-time prediction with UK monetary aggregates in the presence of model uncertainty’, *Journal of Business and Economic Statistics*. Forthcoming. Available as Birkbeck College Discussion Paper No. 0714.
- Garratt, A., Koop, G. & Vahey, S. P. (2008), ‘Forecasting substantial data revisions in the presence of model uncertainty’, *Economic Journal* **118**(530), 1128–1144.
- Garratt, A., Mitchell, J. & Vahey, S. P. (2009), Measuring output gap uncertainty. mimeo, Birkbeck College.
- Genest, C. & Zidek, J. (1986), ‘Combining probability distributions: a critique and an annotated bibliography’, *Statistical Science* **1**, 114–135.
- Geweke, J. & Amisano, G. (2008), Optimal prediction pools. Department of Economics, University of Iowa Working Paper.
- Hall, S. G. & Mitchell, J. (2007), ‘Combining density forecasts’, *International Journal of Forecasting* **23**, 1–13.
- Jore, A. S., Mitchell, J. & Vahey, S. P. (2009), Combining forecast densities from VARs with uncertain instabilities. Working Paper, NIESR, Norges Bank and RBNZ.
- Justiniano, A. & Primiceri, G. E. (2008), ‘The time-varying volatility of macroeconomic fluctuations’, *American Economic Review* **98**(3), 604–41.
- Karagedikli, O., Matheson, T., Smith, C. & Vahey, S. P. (2007), 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., Stock, J. & Watson, M. (2003), ‘A comparison of direct and iterated AR methods for forecasting macroeconomic series h-steps ahead’, *Journal of Econometrics* **135**, 499–526.
- McConway, C. G. . K. J. (1990), ‘Allocating the weights in the linear opinion pool’, *Journal of Forecasting* **9**, 53–73.
- Mitchell, J. & Hall, S. G. (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.
- Mitchell, J. & Wallis, K. F. (2009), Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness. National Institute of Economic and Social Research Discussion Paper No. 320.
- 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., Smith, J. & Hodges, S. (2003), ‘An evaluation of tests of distributional forecasts’, *Journal of Forecasting* **22**, 447–455.
- Raftery, A. E., Gneiting, T., Balabdaoui, F. & Polakowski, M. (2005), ‘Using Bayesian Model Averaging to calibrate forecast ensembles’, *Monthly Weather Review* **133**, 1155–1174.
- Raftery, A. E. & Zheng, Y. (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., Sill, K. & Kryshko, M. (2008), DSGE model-based forecasting of non-modelled variables, Working Papers 08-17, Federal Reserve Bank of Philadelphia.
- Smets, F. & Wouters, R. (2007), ‘Shocks and frictions in us business cycles: A bayesian DSGE approach’, *American Economic Review* **97**(3), 586–606.
- Stock, J. H. & Watson, M. W. (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.
- Wallis, K. F. (2005), 'Combining density and interval forecasts: a modest proposal', *Oxford Bulletin of Economics and Statistics* **67**, 983–994.
- 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:

$$A_t(x) = \left[\eta^{\frac{1}{\mu}} Q_t(x)^{1-\frac{1}{\mu}} + (1-\eta)^{\frac{1}{\mu}} M_t(x)^{1-\frac{1}{\mu}} \right]^{\frac{\mu}{\mu-1}}, \quad (4)$$

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 $h \in [0, 1]$:

$$Q_t(x) = \left[\int_0^1 Q_t(h, x)^{1-\frac{1}{\theta_t}} dh \right]^{\frac{\theta_t}{\theta_t-1}}, \quad (5)$$

where the degree of substitution between domestic intermediate goods, θ_t , evolves according to AR(1) process. Similarly, the composite imported good is a CES aggregate of differentiated import goods indexed $f \in [0, 1]$:

$$M_t(x) = \left[\int_0^1 M_t(f, x)^{1-\frac{1}{\theta_t^*}} df \right]^{\frac{\theta_t^*}{\theta_t^*-1}}, \quad (6)$$

where θ_t^* is the degree of substitution between imported goods, which is also assumed to follow an AR(1) process.

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:

$$T_t(h) = \left[(1-\alpha)^{\frac{1}{\xi}} (Z_t z_t^L l_t(h))^{1-\frac{1}{\xi}} + \alpha^{\frac{1}{\xi}} \bar{K}_t(h)^{1-\frac{1}{\xi}} \right]^{\frac{\xi}{\xi-1}}, \quad (7)$$

where $\alpha \in [0, 1]$ is the capital share and ξ denotes the elasticity of substitution between labour and capital. The variables $l_t(h)$ and $\bar{K}_t(h)$ 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 , whereas z_t^L denotes a temporary (stationary) shock to productivity (or labour utilization). 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

$$\bar{K}_t(h) = u_t(h) K_{t-1}(h), \quad (8)$$

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

$$\gamma_t^u(h) = \phi^{u_1} \left(e^{\phi^{u_2}(u_t(h)-1)} - 1 \right), \quad (9)$$

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:

$$K_t(h) = (1 - \delta) K_{t-1}(h) + \kappa_t(h) K_{t-1}(h), \quad (10)$$

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

$$\begin{aligned} \kappa_t(h) &= \frac{I_t(h)}{K_{t-1}(h)} - \frac{\phi_1^I}{2} \left[\left(\frac{I_t(h)}{K_{t-1}(h)} - \frac{I}{K} \right) \right]^2 \\ &\quad - \frac{\phi_2^I}{2} \left(\frac{I_t(h)}{K_{t-1}(h)} - \frac{I_{t-1}}{K_{t-2}} \right)^2, \end{aligned} \quad (11)$$

where I_t denotes investment and z_t^I is an AR(1) investment shock. The labour input is a CES aggregate of hours supplied by the different households:

$$l_t(h) = \left[\int_0^1 l_t(h, j)^{1 - \frac{1}{\psi_t}} dj \right]^{\frac{\psi_t}{\psi_t - 1}}, \quad (12)$$

where ψ_t is an AR(1) process governing the elasticity of substitution between different types of labour. 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. Nominal price stickiness is modelled by assuming that firms face quadratic costs of adjusting prices,

$$\gamma_t^{P^Q}(h) \equiv \frac{\phi_1^Q}{2} \left[\frac{P_t^Q(h)}{\pi P_{t-1}^Q(h)} - 1 \right]^2 \quad \text{and} \quad \gamma_t^{P^{M^*}}(h) \equiv \frac{\phi_1^{M^*}}{2} \left[\frac{P_t^{M^*}(h)}{\pi P_{t-1}^{M^*}(h)} - 1 \right]^2, \quad (13)$$

in the domestic and foreign market, respectively. 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.

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:

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

where C denotes consumption, l is hours worked and β is the discount factor $0 < \beta < 1$. The consumption preference shock, z_t^u , evolves according to an AR(1) process. The current period utility functions, $u(C_t(j))$ and $v(l_t(j))$, are

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

and

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

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

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

where W_t is the nominal wage rate. The individual flow budget constraint for agent j is:

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

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:

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

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.

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. The central bank sets the interest rate according to a simple instrument rule, which in its log-linearised version takes the form

$$r_t = \lambda^r r_{t-1} + (1 - \lambda^r) \left[\begin{array}{l} \omega_\pi \widehat{\pi}_{t-1} + \omega_y \widehat{gap}_{t-1} + \omega_{rer} \widehat{q}_{t-1} \\ + \omega_{\Delta\pi} (\widehat{\pi}_{t-1} - \widehat{\pi}_{t-2}) + \omega_{\Delta y} \Delta \widehat{gap}_{t-1} \end{array} \right] \quad (20)$$

where π_t is the aggregate inflation rate, and q_t is the real exchange rate. The parameter $\lambda^r \in [0, 1)$ determines the degree of interest rate smoothing. The output gap is measured in deviation from the stochastic productivity trend, the remaining variables are in deviation from their steady-state levels.