Evaluating Predictive Densities of U.S. Output Growth and Inflation in a Large Macroeconomic Data Set*

Barbara Rossi† and Tatevik Sekhposyan‡

February 27, 2013

Abstract

We evaluate conditional predictive densities for U.S. output growth and inflation using a number of commonly used forecasting models that rely on a large number of macroeconomic predictors. More specifically, we evaluate how well conditional predictive densities based on the commonly used normality assumption fit actual realizations out-of-sample. Our focus on predictive densities acknowledges the possibility that, although some predictors can improve or deteriorate point forecasts, they might have the opposite effect on higher moments. We find that normality is rejected for most models in some dimension according to at least one of the tests we use. Interestingly, however, combinations of predictive densities appear to be correctly approximated by a normal density: the simple, equal average when predicting output growth and Bayesian model average when predicting inflation.

Keywords: Predictive Density Evaluation, Structural Change, Output Growth Forecasts, Inflation Forecasts

J.E.L. Codes: C22, C52, C53

*Acknowledgments: We thank the editors, two referees, A. Banerjee, G. Chevillon, D. van Dijk, G. Ganics, E. Granziera, M. Marcellino, C. Schumacher, and participants at the Banque de France and International Institute of Forecasters workshop on "Forecasting the Business Cycle," Deutsche Bundesbank workshop on "Uncertainty and Forecasting in Macroeconomics," the St. Louis Fed Applied Time Series Workshop, and numerous other conferences for useful comments and suggestions. The views expressed in this paper are solely those of the authors and should not be attributed to the Bank of Canada.

†ICREA Research Professor, UPF and Barcelona GSE, CREI, C/ Trias Fargas 25, Barcelona, Spain. Tel.: (+34) 93 542 1655; fax: (+34) 93 542 1860; e-mail: barbara.rossi@upf.edu

‡International Economic Analysis Department, Bank of Canada, 234 Wellington Street, Ottawa, ON, K1A 0G9, Canada. Tel.: (+1) 613-782-7239; fax: (+1) 613-782-7658; e-mail: tsekhposyan@bankofcanada.ca

1 Introduction

Forecasts are traditionally used to evaluate the performance of models. In most cases whether forecasts are good or not is judged mainly based on the models' (median or mean) point forecasts. For example, Stock and Watson (2003) have conducted an extensive evaluation of a large data set of predictors of U.S. output growth and inflation, focusing on point forecasts; Marcellino, Stock and Watson (2003), Banerjee and Marcellino (2006) and Banerjee, Marcellino and Masten (2005) have conducted similarly broad analyses for the Euro area. Rossi and Sekhposyan (2010) have further investigated the stability of point forecasts of output growth and inflation using the same data set. However, it is becoming more and more important to assess the correct specification of uncertainty around models' forecasts. For example, central banks are increasingly concerned about uncertainty around their point forecasts of inflation or unemployment targets, and in particular how well models perform in forecasting a range of future values of important macroeconomic variables.

In this paper we consider models that have been extensively used in the literature for forecasting output growth and inflation (and seemingly doing well according to their point forecasts) and investigate whether their predictive densities are correctly calibrated by the commonly used normal approximation (see Stock and Watson, 2002). We use the Probability Integral Transform (PIT) technique originally introduced by Rosenblatt (1952) and more recently proposed by Diebold, Gunther and Tay (1998) to evaluate the correct specification of predictive densities. Corradi and Swanson (2006b) provide a comprehensive recent overview of tests for predictive density evaluation; Granger and Pesaran (2000) and Garratt, Lee, Pesaran and Shin (2003) further complement the discussion. The difference between this paper and those in the literature is that we operate in a data-rich environment using the extensive data set of Stock and Watson (2003), as well as the wide range of evaluation techniques we use.

The empirical results of this paper are based on several model specifications. Regarding the models, we consider not only predictive densities based on autoregressive distributed lag (ADL) models with several predictors considered one-at-a-time (as in Stock and Watson, 2003), but also forecast combinations. We include predictive density combinations with equal weights or with weights equal to the posterior probabilities of the models. In addition, we consider several estimation techniques: we combine models estimated by OLS as well as via Bayesian shrinkage methods and a posterior simulator algorithm that samples models from the model space with highest posterior probability. Finally, we use methods that pool the

information in various series at the estimation stage as opposed to combining them ex-post, i.e. factor models as well as Bayesian VARs.

We assess the correct specification of predictive densities using several tests. The tests we consider include tests of uniformity, serial correlation and identical distribution. Among the PIT-based tests of uniformity, we consider the histogram-based evaluation technique employed in Diebold, Gunther and Tay (1998) and Diebold, Tay and Wallis (1999), as well as Kolmogorov-Smirnov and Anderson-Darling tests. We also consider tests based on the inverse normal transformation of the PIT, which include the Berkowitz (2001) and Doornik and Hansen (2008) tests. Regarding tests for independence, we consider the Ljung-Box test and a version of Berkowitz's (2001) test for absence of serial correlation (in the PITs).[1] Finally, regarding tests of identical distribution, we consider Andrews' (1993) test of stability applied to the PITs.

Our main empirical findings can be summarized as follows. Overall, the performance of ADL models across the various tests depends crucially on the predictor included in the model. The most interesting result is that pooled predictive densities based on simple averaging as well as Bayesian Model Averaging (BMA) appear to be fairly well calibrated – in particular, the simple model average for one-year-ahead output growth forecasts and the BMA for one-quarter-ahead inflation forecasts. Most of the other models that pool information either at the estimation or at the prediction stage report occasional failings in the correct specification of predictive densities, according to at least one of the tests we consider. Interestingly, the fact that a simple average of several parsimonious ADL models and the BMA have desirable properties in terms of forecasting is a point that has been emphasized many times in the literature in the context of point forecasts (see e.g. Stock and Watson, 2003, Timmermann, 2006 and Wright, 2009), which we find extends to density forecasts when testing the appropriateness of the normal distribution.

In more detail, based on the Kolmogorov-Smirnov and Anderson-Darling tests we find more pervasive evidence against uniformity for predictive densities of inflation relative to output growth, at both short and medium horizons. Similar results hold when assessing the proper calibration of predictive densities in terms of independence: there is more evidence of serial correlation in the PITs of inflation relative to output growth, particularly in the second moment of the PITs. However, there is more evidence of correlation in the PITs of

[1] Note that, throughout this paper, we focus on testing serial correlation in the PITs (as opposed to serial correlation in the forecasts). Serial correlation in the PITs indicates that the pattern of rejection of correct specification is not random over time, and may signal mis-specification in the dynamics of the underlying models.

one-quarter-ahead density forecasts than in one-year-ahead ones. The tests also find some evidence of instabilities in the density forecasts over time, especially at the one-year-ahead horizon; in general, instabilities are more pronounced for output growth than for inflation. Berkowitz's (2001) test confirms the results of no serial correlation in the first moments of the PITs, yet rejects uniformity in a wide set of models of output growth and inflation, particularly at short horizons. However, the normality of the simple average model for output growth and the BMA for inflation is not rejected, with an exception. The exception is that Doornik and Hansen's (2008) test rejects the proper calibration of simple average densities based on non-zero higher (third and fourth) moments of the PITs at the one-quarter-ahead horizon for output growth; it also rejects for the BMA model at the one-year-ahead horizon for inflation.

Overall, under the assumption of normality, predictive densities of simple averaging and BMA models are among the best calibrated despite the target variable we consider. The occasional failings are mainly associated with the higher (greater than first) moments of the PITs when we use the simple average model to forecast inflation at the one-year-ahead forecast horizon and lack of uniformity of the PITs at the one-quarter ahead forecast horizon. Similarly, the BMA performs fairly well for output growth as well, though it fails uniformity for one-quarter-ahead and stability for one-year-ahead forecast horizons.

An analysis similar in spirit to the one considered in this paper is that of Clements and Smith (2000). There are several differences between our work and theirs, however. First, they focus only on forecasting output growth and unemployment, and do not consider inflation forecasts, which is another important variable whose predictive density we are interested in. Furthermore, unlike our paper, they do not consider a large data set of macroeconomic predictors nor a large selection of models, and focus instead on linear and non-linear univariate models and vector autoregressions with selected predictors. Finally, their paper (as well as most papers that evaluate density forecasts, starting from Diebold, Gunther and Tay, 1998) focuses on testing uniformity and uncorrelatedness of the PITs, whereas we also formally test the hypothesis of identical distribution over time.

Our paper is also related to Clark (2011) who, however, focuses on evaluating density forecasts from BVARs, whereas we in addition focus on the linear models and a rich data set of predictors considered by Stock and Watson (2003). Importantly, unlike Clark (2011), our objective is not to improve forecasting models (which in Clark's (2011) paper is accomplished by allowing for stochastic volatility): rather, we consider models that are extensively used in the literature and test whether their density forecasts based on the commonly used normal

approximation are correctly specified.

Our paper is also different from Jore, Mitchell and Vahey (2010) and Manzan and Zerom (2013). Jore, Mitchell and Vahey (2010) combine density forecasts from VARs in the presence of instabilities. We also consider density forecast combinations, but in the presence of large sets of predictors. Finally, note that this paper focuses on testing whether density forecasts of output growth and inflation obtained using a normal distribution are correctly specified, rather than testing which of the competing models' density forecasts are closer to the true but unknown density in the data. The latter can be analyzed using tests proposed by Amisano and Giacomini (2007) and Diks, Panchenko and van Dijk (2011). Importantly, note that we do not undertake an empirical investigation of tests of relative predictive ability in this paper for two reasons: first, our focus is on testing the correct specification of the density forecasts rather than comparing density forecasts; second, a similar analysis has been recently undertaken by Manzan and Zerom (2013), who compare predictive densities of inflation from competing models using selected data from the Stock and Watson (2003) database.[2]

The paper is organized as follows. Section 2 describes the econometric methodology and the tests used in this paper; Section 3 discusses the set of forecasting models, whereas Section 4 describes the data and the empirical results. Section 5 concludes.

2 Econometric Methodology

We are interested in evaluating the $h-$step-ahead predictive density for the scalar variable Y_{t+h}. We assume that the researcher has divided the sample of size $T+h$ observations into an in-sample portion of size R and an out-of-sample portion of size P and obtained a sequence of $h-$step-ahead density forecasts, such that $R+P-1+h=T+h$. Let the sequence of P out-of-sample, estimated conditional predictive densities be denoted by $\left\{\widehat{\phi}_{t+h}\left(Y_{t+h}|\Im_t\right)\right\}_{t=R}^{T}$, where $\Im_t$ is the information set at time t. We obtain the conditional predictive densities under the normality assumption by estimating the parameters in the conditional moments using a rolling window procedure. Thus, $\widehat{\phi}_{t+h}$ denotes the probability density function (PDF) of a normal distribution where the parameters are re-estimated at each $t=R,...,T$ over a window of R observations including data indexed $t-R+1$ to t. The rolling window estimation procedure is more robust to breaks in the conditional moments of the predictive densities and has a better chance to result in properly calibrated densities –

[2]Other related papers include those considering measures of uncertainty such as Guidolin and Timmermann (2006).

see Clark (2011) and Jore, Mitchell and Vahey (2010).

We test whether the realized values $\{Y_{t+h}\}_{t=R}^{T}$ are generated by $\left\{\hat{\phi}_{t+h}\left(Y_{t+h}|\Im_t\right)\right\}_{t=R}^{T}$ using the Probability Integral Transform (PIT) approach suggested by Diebold, Gunther and Tay (1998). For a given probability density function $\widehat{\phi}_{t+h}$, the PIT is the corresponding cumulative density function (CDF) evaluated at the realization Y_{t+h}:

$$z_{t+h} = \int_{-\infty}^{Y_{t+h}} \hat{\phi}_{t+h}\left(u|\Im_t\right) du \equiv \widehat{\Phi}_{t+h}\left(Y_{t+h}|\Im_t\right) \tag{1}$$

According to Diebold, Gunther and Tay (1998), if the proposed predictive density is consistent with the true predictive density then, for $h = 1$, the density of $\{z_{t+h}\}_{t=R}^{T}$ is an independent and identically distributed (iid) Uniform (0,1) and its cumulative distribution function is the 45° line. When $h > 1$, then independence is violated by construction, even if models are correctly specified, since serial correlation of order $(h-1)$ is built in by construction in the multi-step ahead density forecasts. One recommendation given in Diebold, Gunther and Tay (1998) and Clements and Smith (2000), among others, is to split the sample into independent sub-samples where the PITs are at least h-periods apart. In this case inference on the proper calibration of the predictive densities can be done separately in each of the sub-samples, or jointly via Bonferroni bounds.

In what follows, we consider several tests, each of which focuses on different properties that correctly specified PITs should satisfy. In choosing which test to implement, we follow Mitchell and Wallis (2011) and focus on the Ljung-Box (LB), Kolmogorov-Smirnov (KS), Anderson-Darling (AD), Berkowitz (2001) and Doornik and Hansen (2008) tests. The first test aims only at detecting the absence of serial correlation in the PITs; the rest of the tests aim at detecting violations of uniformity (at times joint with independence); in particular, the last two tests operate not on the PITs directly, but rather on the inverse normal transformation of the PITs. In addition, we implement Andrews' (1993) QLR test to evaluate the stability (i.e. identical distribution) of predictive densities which should be satisfied if they are properly calibrated.[3]

It is important to note that these tests have different properties. For example, both Mitchell and Wallis (2011) and Noceti, Smith and Hodges (2003) document the power advantage of the AD test over the KS test in Monte Carlo simulation exercises. Berkowitz

[3]Note that none of the tests considered here account for parameter uncertainty. As discussed in Berkowitz (2001) and the references therein, parameter estimation error is empirically of second-order importance in the presence of model mis-specification. For discussion of tests that take into account parameter estimation uncertainty, see Corradi and Swanson (2006b).

(2001), on the other hand, suggests that the tests of proper calibration based on the inverse normal of the PITs (such as those proposed by Berkowitz, 2001) are more powerful than the tests of uniformity applied directly to the PITs, at least in finite samples. In what follows, we discuss in detail the characteristics of each of the tests we implement.

(a) Tests on the PIT

I. Diebold, Gunther and Tay (1998) test. Diebold, Gunther and Tay (1998) mainly rely on a graphical assessment of uniformity and independence properties that characterize PITs of correctly specified predictive distributions. Following Diebold, Gunther and Tay (1998), we test the uniformity of the empirical distribution function of the PITs (i.e. the histogram of the PITs); independence is assessed by reporting the autocorrelation function of various powers of the PITs. More in detail, in order to statistically assess the uniformity of the PITs, we follow Diebold, Gunther and Tay (1998) in deriving confidence intervals for the number of observations falling into any bin; under the maintained assumption of independence, the latter follows a binomial distribution. We divide the unit interval into $n_b = 5$ equally sized bins and depict the fraction of PITs falling into each bin. If the PITs are indeed iid uniform, then each bin would contain $\hat{p} = 100/n_b\% = 20\%$ of the PITs. We construct the 2.5th and 97.5th percentiles of the distribution of $\hat{p}$ by using a normal approximation: $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/P}$. The usefulness of this approach is that, when the PITs are not uniform (i.e. the empirical distribution function of the PIT fails to have a rectangular shape), the shape of the histogram sheds light on the reasons behind the failure of the model.

II. Tests of uniformity

We test whether the PIT is uniform using the Kolmogorov-Smirnov and Anderson-Darling tests. The latter measures the difference between the empirical distribution of the PITs, $\widehat{\Phi}_{t+h}\left(y_{t+h}|\Im_t\right)$, and the cumulative distribution of a uniform, $r \in (0,1)$ (i.e., the 45° line). Anderson-Darling test is a special type of Cramér-von Mises test, which puts more weight on the deviations in the tails of the empirical distribution, as opposed to weighting all its points equally. We implement these tests following Kroese, Taimre and Botev (2011, Chapter 8). Let $z_j^\dagger$ denote the values of z_{t+h} in ascending order. The test statistics are provided below:

i. Kolmogorov-Smirnov (KS, Kolmogorov, 1933, and Smirnov, 1948)

$$KS = \sqrt{P} \underset{j=1,\ldots,P}{max} max \left\{ |z_j^\dagger - j/P|, |z_j^\dagger - (j-1)/P| \right\}; \tag{2}$$

ii. Anderson-Darling (AD, Anderson and Darling, 1952, 1954)

$$AD = -P - \frac{1}{P}\sum_{j=1}^{P}(2j-1)ln(z_j^{\dagger}(1-z_{P+1-j}^{\dagger})) \tag{3}$$

Both KS, as well as AD tests have non-standard asymptotic distributions. We obtain their critical values based on the approximations detailed in Kroese, Taimre and Botev (2011, Chapter 8).[4]

III. Test for independence (Ljung-Box)

We test for independence in the first and second central moments of the PITs via the Ljung-Box test of serial correlation.[5] The test statistic is

$$Q = P(P+2)\sum_{l=1}^{\widetilde{L}}\left(\frac{\rho(l)^2}{P-l}\right), \tag{4}$$

where $\rho(L)$ is the serial correlation coefficient at lag l of either the demeaned PITs or their square. We implement this test with a maximum lag length $\widetilde{L}$ equal to 4, given the quarterly nature of our data. The p-values are based on an asymptotic $\chi^2(l)$ distribution, which approximates the distribution well even in moderate sample sizes (cfr. Hayashi, 2000, p. 144).

IV. Tests of identical distribution

To complement the empirical evidence, we also consider tests for identical distribution. If z_{t+h} were identically distributed over time, then its (non-central) moments would be constant over time. We consider empirical evidence on the time variation in the PITs by reporting Andrews' (1993) QLR test for structural breaks. The test has been typically used in the forecasting literature to judge whether predictors' Granger-causality is stable over time: see Stock and Watson (2003). Here, we are concerned about whether the distribution of the PITs has changed over time, and thus we test whether $\alpha_{1,t}$ and $\alpha_{2,t}$ are constant in each of

[4] Alternatively, one could simulate their critical values as in Mitchell and Wallis (2011).

[5] We report the results for the first and second (rather than on higher) moments only due to space constraints.

the following regressions:[6]

$$z_{t+h} = \alpha_{1,t} + \varepsilon_{1,t+h} \tag{5}$$

$$z^2_{t+h} = \alpha_{2,t} + \varepsilon_{2,t+h}. \tag{6}$$

(b) Tests on the Inverse Normal of the PIT

Berkowitz (2001, Proposition 1) shows that if the PIT is iid U(0,1), then the inverse standard normal transformation of the PIT is an iid Normal (0,1). Let the inverse standard normal transformation of the PIT be denoted by $\widetilde{z}_{t+h}$, where $\widetilde{z}_{t+h} \equiv \widetilde{\Phi}^{-1}(z_{t+h})$ and $\widetilde{\Phi}(.)$ is the standard normal CDF. We implement two tests on this transformation.

I. Berkowitz's (2001) test. Berkowitz (2001) proposes a joint test for zero mean, unit variance, and independence in $\widetilde{z}_{t+h}$, against an autoregressive alternative with a mean and a variance possibly different from 0 and 1, respectively. That is, we jointly test whether $\mu = 0$, $\sigma = 1$ and $\rho = 0$ in the regression:

$$\widetilde{z}_{t+h} - \mu = \rho(\widetilde{z}_t - \mu) + \varepsilon_{t+h}, \tag{7}$$

where $\varepsilon_{t+h} \sim (0, \sigma^2)$.[7] This test is implemented as a likelihood ratio (LR) test, which under the null hypothesis described above, has an asymptotic $\chi^2(3)$ distribution. One could also test a subset of the hypotheses in this setting, for example test independence ($\rho = 0$), which has an asymptotic distribution equal to a $\chi^2(1)$ under the null hypothesis. The difference between this test and the ones under the PIT framework is that Berkowitz (2001) is a joint test of independence and normality for the inverse normal transformation of the PIT. According to Berkowitz (2001), the advantage of tests based on the inverse normal transformation of the PITs is that they are more powerful than tests of uniformity applied directly to the PITs, at least in small samples; the limitation is that they detect violations of normality only through the first two, and not higher, moments, whereas PIT-based tests can detect any departure from uniformity.

II. Doornik and Hansen's (2008) test. Doornik and Hansen (2008) propose to test the normality of $\widetilde{z}_{t+h}$ using a test on skewness and kurtosis which has good small sample properties. The test is based on the sum of the squares of transformed measures of skewness

[6] While, for simplicity, we use Andrews' (1993) test for parameter stability on the PIT, a better approach would be to use the test for stability of the distribution proposed by Rossi and Sekhposyan (2012), as the latter is specifically designed for densities and could also be used to take into account parameter estimation error.

[7] Eq. 7 could be generalized to include higher-order dependence.

and kurtosis, and has a $\chi^2(2)$ asymptotic distribution under the null of iid normality (i.e. absence of skewness and kurtosis).

3 Forecasting Models

All the models we consider are estimated using the Stock and Watson (2003) database collected at the quarterly frequency and updated up to January 2011. These variables are asset prices, measures of real economic activity, wages and prices, and money. We follow Stock and Watson (2003) and transform the data to eliminate stochastic or deterministic trends, as well as seasonality. For example, all the variables that represent rates are considered in levels, while the rest are considered in natural logarithmic differences. For a detailed description of the variables we consider and their respective transformations, see Table 1. The variables are in percentage points, and the growth rates have been annualized. The earliest starting point of the sample that we consider is January 1959, although several series have a later starting date due to data availability constraints. We use a fixed rolling window estimation scheme with a window size of 40 observations. For simplicity, when describing the models below, we omit the time-subscript that would be appropriate given the time-varying nature of the parameters introduced by the rolling window estimation.

INSERT TABLE 1 HERE

We consider an ADL model, where individual predictors are used one-at-a-time, as well as models that pool information across series, such as BMAs, BVARs and factor models. In what follows, we describe these models and their implied PITs.

3.1 Autoregressive Distributed Lag (ADL) Models

We consider forecasting quarterly output growth and inflation h-periods into the future using lags of one predictor at a time in addition to the lagged dependent variable. The forecasting model is:

$$Y_{t+h}^h = \beta_{k,0} + \beta_{k,1}(L) X_{t,k} + \beta_{k,2}(L) Y_t + u_{t+h},\ t = 1, ..., T \tag{8}$$

where the dependent variable is either $Y_{t+h}^h = (400/h)\ln(RGDP_{t+h}/RGDP_t)$ or $Y_{t+h}^h = 400/h\ln(PGDP_{t+h}/PGDP_t) - 400\ln(PGDP_t/PGDP_{t-1})$, where $RGDP_{t+h}$ and $PGDP_{t+h}$ are the real GDP and GDP deflator, respectively. X_t is the $1 \times K$ vector of explanatory variables in Stock and Watson's (2003) database, and $X_{t,k}$ denotes the k-th variable,

for $k = 1, ..., K$. Note that the total number of individual economic variables considered in our application is $K = 32$.[8] Y_t is either the period t output growth, that is $Y_t = 400 \ln(RGDP_t/RGDP_{t-1})$, or the period t change in inflation, that is $Y_t = 400 \ln(PGDP_t/PGDP_{t-1}) - 400 \ln(PGDP_{t-1}/PGDP_{t-2})$.[9] Further, the error term u_{t+h} is assumed to be distributed normally, $N(0, \sigma^2)$. We consider $h = 1, 4$ corresponding to one-quarter ahead and one-year ahead forecast horizons. $\beta_1(L) = \sum_{j=0}^{p} \beta_{1j} L^j$ and $\beta_2(L) = \sum_{j=0}^{q} \beta_{2j} L^j$, where L is the lag operator. We estimate the number of lags (p and q) recursively by BIC, first selecting the lag length for the autoregressive component, then augmenting with an optimal lag length for the additional predictor. The PITs at a given time period $t + h$ are: $\Phi_{t+h}(Y_{t+h}^h | (\hat{\beta}_0 + \hat{\beta}_1(L) X_{t,k} + \hat{\beta}_2(L) Y_t), \hat{\sigma}^2)$, where $\hat{\cdot}$ indicates OLS estimates of the model's parameters, while Φ_{t+h} is the conditional CDF of the proposed normal distribution. To estimate $\hat{\sigma}^2$, we use HAC-robust variance estimates (Newey and West, 1987).[10]

As a particular case, we consider the autoregressive model, where we use only the lagged dependent variable to forecast output growth and inflation. The PIT for the autoregressive model is $\Phi_{t+h}(Y_{t+h}^h | (\hat{\beta}_0 + \hat{\beta}_2(L) Y_t), \hat{\sigma}^2)$, where the predictive distribution is again assumed to be normal and the conditional moments are obtained similarly to those of the ADL models.

3.2 Pooled Models

We consider several models.

(i) Simple Average Model. The first pooling strategy we consider is the simple model average which has been shown to perform well for point forecasts by Stock and Watson (2003, 2004).[11] We follow Mitchell and Wallis (2011), and consider the predictive distribution of the combined model. More in detail, we estimate the ADL models in eq. (8) for all the regressors (one-at-a-time), i.e. for $k = 1, ..., K$, and consider linear combinations of their PITs, where each PIT is weighted with an equal weight $(1/K)$. The PIT associated with the equal-weight pooled predictive density is defined as (see Jore, Mitchell and Vahey, 2010, eq. 1):

$$\Phi_{t+h}^c = \frac{1}{K} \sum_{k=1}^{K} \Phi_{t+h}(Y_{t+h} | \hat{\beta}_{k,0} + \hat{\beta}_{k,1}(L) X_{t,k} + \hat{\beta}_{k,2}(L) Y_t, \hat{\sigma}_k{}^2), \tag{9}$$

where the k subscripts in the conditional moments indicate that parameters correspond to

[8] The datasets for output growth include historical data for inflation, but not output growth (and vice versa), as the lagged dependent variable is automatically included in eq. (8).

[9] Note that, like Stock and Watson's (2003) approach, this relies on the assumption that inflation is I(2).

[10] The truncation parameter used in the HAC estimate is $R^{1/4}$.

[11] See Timmermann (2006) for a review of forecast combination.

the k-th ADL regression.[12]

(ii) Bayesian Model Averaging (BMA). The second averaging method we consider is the Bayesian Model Average, which also pools from the set of simple models, yet assigns weights that are proportional to the models' posterior probabilities. BMA puts more weight on more likely models as opposed to putting equal weight on all the models. We consider two variants of BMA models following Wright (2009). Note, however, that Wright (2009) is concerned with model averaging in point forecasts whereas we are interested in BMA for density forecasts.

- *BMA-OLS.* The first version is very similar to the simple model average (eq. 9) as it uses the OLS estimates of the respective model's parameters. It is different however from the simple model average since it has time-varying weights $P_t(M_k|D_t)$, which represent the posterior probability of model k denoted by M_k, given the data $D_t = \{Y_t, X_t, Y_{t-1}, X_{t-1}, ..., Y_{t-R}, X_{t-R}\}$. The PIT in this case is:

$$\Phi_{t+h}^{BMA-OLS} = \sum_{k=1}^{K} P_t(M_k|D_t)\Phi_{t+h}(Y_{t+h}|(\hat{\beta}_{k,0} + \hat{\beta}_{k,1}(L)X_{t,k} + \hat{\beta}_{k,2}(L)Y_t), (\hat{\sigma}_k)^2) \tag{10}$$

-BMA. The second version of BMA we consider is the full Bayesian version, where the estimated parameters are not the OLS counterparts (in the Bayesian framework this would be equivalent to obtaining coefficients under a flat prior), but rather they are posterior estimates and, thus, are influenced by the choice of the prior distribution. Let $\tilde{}$ indicate estimates associated with the fully Bayesian estimation. In this case the PIT is the weighted average of the cumulative predictive densities, denoted by $\tilde{\Phi}_{t+h}$, using weights that are the posterior probabilities of their respective models:

$$\Phi_{t+h}^{BMA} = \sum_{k=1}^{K} P_t(M_k|D_t)\tilde{\Phi}_{t+h}(Y_{t+h}|D_t, M_k), \tag{11}$$

where M_k denotes the k-th model and $P_t(M_k|D_t)$ is the posterior probability of the k-th model given the data D_t.

We follow Wright (2009) and apply a g-prior for $\tilde{\beta}_k = [\tilde{\beta}_{k,0}\ \tilde{\beta}_{k,11}\ ...\ \tilde{\beta}_{k,1p}\ \tilde{\beta}_{k,21}\ ...\ \tilde{\beta}_{k,2q}]'$. More specifically, let $\tilde{X}_k$ denote the $T \times (q + p + 1)$ matrix of explanatory variables and Y^h as the $T \times 1$ dependent variable, then

$$\tilde{\beta}_k|\tilde{h}_k \backsim N(\bar{\beta}_k, \tilde{h}_k^{-1}[g\tilde{X}_k'\tilde{X}_k]^{-1}), \tag{12}$$

[12]Note that we do not consider the simple AR model in the model combinations.

where $\tilde{h}_k = \tilde{\sigma}_k{}^{-2}$ is the precision parameter. We follow Koop (2003, Chapter 3) and assume a Gamma prior distribution for the precision parameter

$$\tilde{h}_k \backsim G(\bar{s}_k^{-2}, \bar{\nu}). \tag{13}$$

We set $\bar{\nu} = 0$ which creates an uninformative prior for the precision (i.e. the variance of the regression equation). This is appropriate since the precision parameter is common to all models. As in Wright (2009), we assume $g = 1$, which puts equal weight on the prior and the data in the posterior density of regression coefficients. To further parameterize the prior, we need values for $\bar{\beta}_k = [\bar{\beta}_{k,0}\ \bar{\beta}_{k,10}\ ...\ \bar{\beta}_{k,1p}\ \bar{\beta}_{k,20}\ ...\ \bar{\beta}_{k,2q}]'$ and $\bar{s}_k^2$. $\bar{\beta}_0^k$ and $\bar{\beta}_{10}^k$ are set to their pre-estimation sample values obtained from autoregressive (of order 1) models of inflation and output growth estimated over 1947:Q1-1958:Q4, while the remaining coefficients are centered around zero.[13]

We obtain the posterior distributions by adapting Koop (2003, Chapter 3 and Chapter 11) to our prior distributions:

$$\tilde{\beta}_k, \tilde{h}_k | D_t \sim NG(\underline{\beta}, \underline{V}, \underline{s}^{-2}, \underline{\nu}), \tag{14}$$

where NG(.) denotes the Normal-Gamma distribution. Let $\hat{\beta}$ denote the OLS estimate of the regression coefficients and $P_{\tilde{X}_k} = I_T - \tilde{X}_k(\tilde{X}_k'\tilde{X}_k)^{-1}\tilde{X}_k'$; then

$$\underline{V} = [(1+g)\tilde{X}_k'\tilde{X}_k]^{-1} \tag{15}$$

$$\underline{\beta} = \frac{\hat{\beta}_k}{1+g} + \frac{\bar{\beta}_k g}{1+g} \tag{16}$$

$$\underline{\nu} = T \tag{17}$$

$$\underline{s}^2 = \underline{\nu}^{-1}\left[\frac{1}{1+g}Y^{h\prime}P_{\tilde{X}_k}Y^h + \frac{g}{1+g}(Y^h - \tilde{X}_k\bar{\beta}_k)'(Y^h - \tilde{X}_k\bar{\beta}_k)\right] \tag{18}$$

Furthermore, in this context both the predictive density as well as the posterior model density have analytical solutions. The predictive density is given by

$$Y_{t+h}^h | D_t, M_k \sim t(\tilde{X}_t\underline{\beta}, \underline{s}^2[I_T + \tilde{X}_t\underline{V}\tilde{X}_t'], \underline{\nu}), \tag{19}$$

[13]It turns out that by setting $\bar{\nu} = 0$, we yield the specific value of $\bar{s}_k^{-2}$, which is irrelevant for further calculations. The mean of the gamma distribution is defined by $\bar{s}_k^{-2}\bar{\nu}$, while the variance is $\bar{s}_k^{-2}\bar{\nu}^2$, which both become zero when $\bar{\nu} = 0$. This would be equivalent of having no prior (or having an uninformative prior) for the precision despite the specific value of $\bar{s}_k^{-2}$.

For the degrees of freedom implied by our rolling sample size of 40, the t-distribution is similar to a normal distribution. On the other hand, under the assumption that all the models are a priori equally likely, the model's posterior distribution becomes

$$p(M_k|D_t) = \frac{p(Y^h|M_k)}{\sum_{j=1}^{K} p(Y^h|M_j)} \tag{20}$$

and the marginal likelihood $p(Y^h|M_k)$ is described as being proportional to

$$p(Y^h|M_k) \propto \left(\frac{g}{1+g}\right)^{\frac{(1+p+q)}{2}} [\underline{\nu s}^2]^{-\frac{T-1}{2}}. \tag{21}$$

Note that when $g = 0$, both the BMA-OLS and the BMA models reduce to the simple model average, **as** $g = 0$ is equivalent to estimating parameters under a flat prior and assigning each individual model a weight equal to $1/K$. In addition, the lag selection is important. When considering the ADL models or the simple model average, p and q (the lag length) are selected recursively via BIC. We keep p and q fixed at their recursively selected levels for both the BMA-OLS as well as the BMA specifications. Furthermore, as noted in Wright (2009), the analytical results presented in this section work under the assumption of strict exogeneity of the regressors and do not allow for serial correlation in the error terms, which is very important given our multi-step forecasts. One could allow for serial correlation, but this would come at a cost of not being able to derive analytical solutions for the predictive densities and models' posterior probabilities. The latter would require a simulation, which could be numerically intensive. Since the point forecasting literature has shown that models could have good forecasting properties even if their theoretical assumptions are not fully satisfied, we proceed under the assumption that the BMA could still perform well in terms of predictive densities.

-BMA-MC3. The last model averaging technique we consider is the Markov Chain Monte Carlo Model Composition (MC3). The theoretical framework of the BMA-MC3 is very similar to that of the BMA, except that the former is a posterior simulation algorithm which allows to consider a multiplicity of models at a lower computational cost: in fact, it allows all regressors to enter the right hand side of the regression model (and not just the autoregressive lags and lags of only one additional economic variable). That is, MC3 is an algorithm that could help the researcher sample from the model space by concentrating on the regions where the models' posterior probabilities are high – see Koop (2003, Chapter 11) for the algorithm, which we extend to pooling models' predictive densities. More specifically, the algorithm is:

- Start with a model M^0. In our case, we start with the autoregressive model with the lag length of q and one additional explanatory variable.[14]

- At step s, $s = \{1, 2, ..., S\}$, consider a new candidate model M^*, which is drawn randomly with equal probability from set of models that include: (i) the current model M^{s-1}; (ii) all models that add one additional explanatory variable to the current model M^{s-1}; (iii) all models that delete one explanatory variable from the current model M^{s-1}.

- We accept the candidate model with probability:

$$\alpha(M^{s-1}, M^*) = min\left[\frac{p(Y|M^*)}{p(Y|M^{s-1})}, 1\right] \tag{22}$$

- We save $P_t(M_k|D_t)$ and $\tilde{\Phi}_{t+h}(Y_{t+h}|D_t, M_k)$ for accepted models.

Let $S = 10,000$ be the total number of draws, while $\bar{S} = 1,000$ denotes the number of burn-in draws.[15] The pooled predictive density is:

$$\Phi_{t+h}^{MC3} = \sum_{s=\bar{S}+1}^{S} P_t(M_s|D_t)\tilde{\Phi}_{t+h}(Y_{t+h}|D_t, M_s). \tag{23}$$

3.3 Models with Principal Components

Next, we consider a variant of the ADL model, eq. (8), where instead of considering each individual regressor one-by-one, we consider one model augmented with factors extracted from the set of all regressors. More in detail, we estimate a static factor model:[16]

$$Y_{t+h}^h = \beta_0 + \boldsymbol{\gamma}\hat{F}_t + \beta_2(L)Y_t + u_{t+h}^h,\ t = 1, ..., T, \tag{24}$$

where $\hat{F}_t$ is the $(1 \times m)$ vector of estimated first m principal components of the K variables we consider in this paper. We recursively select the number of factors m over each rolling window R such that the total number of factors explain at least 60% of the variation contained in

[14] Given our large database we do not consider the lags of economic variables, since that would make the model space, which is already large, even larger, and less feasible to simulate.

[15] Draws discarded to minimize the effect of the starting point in the simulation.

[16] The static factor model could, in principle, be extended to a dynamic factor model, although, as Bai and Ng (2007) note, there is little gain to be expected from moving from static to dynamic factor models from a forecasting standpoint.

the K macroeconomic data series. This results in 2-3 factors for output growth and inflation at different estimation periods.[17] The remaining definitions from the ADL models carry forward to this case: the PIT is $\Phi_{t+h}(Y^h_{t+h}|(\hat{\beta}_0 + \gamma\hat{F}_t + \hat{\beta}_2\,(L)\,Y_t), \hat{\sigma}^2)$, where $\hat{\cdot}$ indicates OLS estimates of the model parameters, while Φ_{t+h} is the conditional CDF of the proposed normal distribution, and σ^2 is estimated by HAC.

3.4 Bayesian Vector Autoregressions

Finally, we consider a large scale Bayesian vector autoregression (BVAR) to model the joint dynamics of all the variables simultaneously. Our BVAR(l) specification is:

$$\mathcal{Y}^h_{t+h} = C + B(L)\mathcal{Y}_t + U_{t+h}, \tag{25}$$

where $\mathcal{Y}^h_{t+h} = [Y^h_{1,t+h}\ , Y^h_{2,t+h}\ ,\ X^h_{1,t+h}\ ,\ldots\ X^h_{k,t+h}, \ldots X^h_{K,t+h}]'$, $Y^h_{1,\ t+h}$ and $Y^h_{2,\ t+h}$ are the h-step-ahead variables for output growth and inflation defined as in eq. (8) and $X^h_{k,t+h} = (400/h)\ln(X_{t+h}/X_t)$, $\mathcal{Y}_t = [Y_{1,t},\ Y_{2,t},\ X_{1,t},\ \ldots\ X_{k,t},\ \ldots\ X_{K,t}]'$, $Y_{1,\ t}$ and $Y_{2,\ t}$ are the output growth and inflation in period t, U_{t+h} is a $(K+2)\times 1$ error term, $U_{t+h} \sim N(0, \Sigma_u)$ and $B(L) = \sum_{j=0}^{l} B_l L^j$, where L is the lag operator and l is selected recursively by BIC. We assume that Σ_u is proxied by the sample variances of the respective series $\Sigma_u = diag(\hat{\sigma}^2_1,\ \ldots, \hat{\sigma}^2_{K+2})$ over the respective rolling estimation windows.

Let $B = [C\ B_1\ \ldots\ B_l]'$ and $\beta = vec(B)$. We impose a conditional prior on β

$$\beta|\Sigma_u \sim N(vec(\bar{B}), \Sigma_u \otimes \bar{\Omega}\lambda^2).$$

We parameterize the prior such that it centers the regression coefficients around zero ($\bar{\beta} = 0$) reflecting our prior belief on the mean reverting nature of the variables in the VAR (all series except the rates are in first differences.) Further, $\bar{\Omega}$ is parameterized such that the coefficients on the lagged variables are independent of each other and the covariance matrix for each lagged coefficient is parameterized as:

$$Var((B_l)_{ij}) = \begin{cases} \frac{\lambda^2}{l^2}, & j = i \\ \frac{\lambda^2}{l^2}\frac{\hat{\sigma}^2_i}{\hat{\sigma}^2_j}, & \text{otherwise} \end{cases} \tag{26}$$

[17] Note that the datasets for output growth includes historical data for inflation, but not output growth and vice versa. We also considered the IC_{p1} criterion of Bai and Ng (2002), but for our data set it chooses a very large number of factors: it chooses 10 when the maximum number of factors is allowed to be 10, for example.

where the $\hat{\sigma}_i$ is re-estimated over each rolling window.

The prior variance on the constant is simply λ^2. Given the quarterly nature of the data we set $\lambda = 0.2$ as recommended in Sims and Zha (1998). Given the results in Giannone, Lenza and Primiceri (2012), under the assumption that the variance matrix of the residuals Σ_u is known, the conditional predictive density for an individual variable such as output growth and inflation (separately) can be written as:

$$\Phi_{t+h}^{BVAR} = \Phi_{t+h}(\mathcal{Y}_{i,t+h}^h | \mathcal{Y}_t' \hat{B}_i, \mathcal{Y}_t' \hat{V}_i \mathcal{Y}_t + \hat{\sigma}_i), i = 1, ..., K+2 \tag{27}$$

where $\Phi_{t+h}(.)$ is the CDF of the normal distribution, $\mathcal{Y} = [1_k \; \mathcal{Y}_{l+1} \; ... \; \mathcal{Y}_T]$, $\mathcal{Y}_i^h = [\mathcal{Y}_{i,l+1}^h \; ... \; \mathcal{Y}_{i,T}^h]$ and

$$\hat{B}_i = (\mathcal{Y}\mathcal{Y}' + (\bar{\Omega}\lambda^2)^{-1})^{-1}(\mathcal{Y}\mathcal{Y}_i^{h\prime} + (\bar{\Omega}\lambda^2)^{-1}\bar{B}_i) \tag{28}$$

$$\hat{V}_i = \sigma_i \otimes (\mathcal{Y}\mathcal{Y}' + (\bar{\Omega}\lambda^2)^{-1})^{-1}. \tag{29}$$

Note that the estimator of the variance ($\hat{V}_i$) is not HAC; rather, it relies on iid assumptions to obtain a simple analytical (normal) solution for the predictive density.

4 Empirical Evidence

This section discusses the empirical evidence. We start by considering tests of uniformity for both medium and short horizon forecasts, one-year- and one-quarter-ahead, respectively. The PIT-based tests of uniformity include: Diebold, Gunther and Tay (1998), Kolmogorov-Smirnov, and Anderson-Darling. Then we discuss tests of independence; finally, we provide tests for identical distribution (instabilities). We conclude by considering tests for the correct specification based on the inverse normal transform of the PITs.

To preview our results, we find that there is more evidence against lack of uniformity for density forecasts of inflation than for output growth, at both short and medium horizons. Our main result is that the best calibrated predictive densities (in terms of correct calibration by a normal density) are density combinations, in particular the simple averaging for one-year-ahead output growth forecasts, and Bayesian model averaging for one-quarter-ahead inflation forecasts. The autoregressive model, the factor model and a variant of Bayesian model average constructed with OLS estimates perform fairly well in terms of correct specification for output growth at the one-quarter-ahead horizon as well, though the correct specification of normal density forecasts fails for all other models according to at least one of the tests we

consider.

Regarding correlation, in general, forecast densities are fairly well calibrated for GDP growth, with occasional exceptions, but less so for inflation; in addition, there is more evidence of correlation in the PITs of one-quarter-ahead forecasts than in one-year-ahead ones. Several versions of model averaging, as well as the factor model, perform fairly well, though the factor model and the simple average show evidence of serial correlation in the second moments of the PITs in the case of inflation.

The tests also find some evidence of instabilities in the density forecasts over time, especially for one-year-ahead forecasts, and more so for output growth than inflation.

Overall, across the various tests we consider, the performance of the ADL model depends crucially on the predictor, the forecast horizon and the target variable.[18]

4.1 Test of Uniformity

Figures 1 to 2 report results based on the Diebold, Gunther and Tay (1998) test for one-quarter-ahead density forecasts. Figure 1 focuses on forecasts of output growth whereas Figure 2 focuses on inflation. In each figure, the pictures in Panel (A) report the empirical distribution of the PIT for ADL models, eq. (8), with selected regressors (such as the lagged dependent variable, or autoregressive (AR) model, reported on the top, left side; the spread; unemployment; and money (M1)); Panel (B) instead reports results on the PITs for models combining large data sets: the equal combinations of density forecasts across the ADL models, eq. (9), labeled "Simple Average"; the BMA model with OLS weights, eq. (10), labeled "BMA-OLS"; the BMA model with posterior weights, eq. (11), labeled "BMA"; the BMA model with MC3, eq. (23), labeled "BMA-MC3"; the factor model, eq. (24), labeled "Factor"; and the BVAR model, eq. (27), labeled "BVAR". In addition to the empirical distribution function of the PIT, the pictures report 95% confidence bands for the null hypothesis of iid uniformity.

Figure 1(A,B) shows that, when forecasting output growth, Diebold, Gunther and Tay's (1998) test rejects the hypothesis of normality (under the maintained assumption of independence) for the ADL model with the unemployment rate, the BMA model, as well as

[18] For instance, models based on nominal interest rates appear to result in correctly calibrated densities for output growth, but not for inflation, except for the Berkowitz (2001) test at short horizons. The simple autoregressive model appears to be well calibrated benchmark at the one-year-ahead forecast horizon for output growth, while it fails according to the Berkowitz's (2001) test for both short-term forecasts of output growth and inflation, as well as in the dimension of second and higher moments of the PITs for the one-year-ahead inflation.

BMA-MC3 and the BVAR. The histograms of the PITs of the BMA-MC3 and the BVAR suggest that too many realizations fall in the middle of the distribution relative to what would have been expected if the PITs were iid uniform.

Figure 2(A,B) instead shows results for density forecasts of inflation at the one-quarter-ahead horizon. In this case the test does not reject uniformity for the ADL models we display (see Panel (A)). Density combinations (reported in Panel (B)) appear to be well calibrated, with the exception of the BMA-MC3 and the BVAR models: they again overestimate the realizations in the middle, but less severely than in the case of output growth.

INSERT FIGURES 1 AND 2 HERE

For the same models, Tables 2 and 3 provide results for the Kolmogorov-Smirnov (labeled "KS") and the Anderson-Darling (labeled "AD") tests of uniformity of the PITs, which test the correct specification of the predictive densities, again under the assumption of independence. Table 2 reports results for short-horizon predictive densities. The left panel in Table 2 shows that, when predicting GDP growth, the KS test mostly favors correct specification across the models, while the AD test finds strong evidence of mis-specification for most of the predictors, with the exception of various nominal interest rates, industrial production, employment and some measures of money. In addition, the tests (in particular the AD test) detect mis-specification in all the BMA model specifications as well as the BVAR.[19] The simple average and the factor models are, however, correctly specified. Note that, relative to the results reported in Figure 1, the AD test is slightly more powerful in detecting mis-specification in several models (e.g. the BMA-OLS and BMA). The right panel in Table 2 shows that, when predicting inflation, most of the predictors and models result in mis-specified densities according to the AD test (although not according to the KS test); only the simple average and the BMA models are correctly specified according to both the KS and AD tests, while the densities of the factor and BMA-OLS models are mis-specified according to the AD test. Note that oftentimes the KS and AD tests reach opposite conclusions: the discrepancies between the tests are most likely due to the higher power of the AD test relative to the KS test, especially in the tails of the distributions, which we alluded to in Section 2. In the case of inflation, the AD test finds more empirical evidence of mis-specification than in the case for output growth. In addition, it also finds more evidence of mis-specification than Diebold, Gunther and Tay's (1998) test, especially for several ADL models. Overall, equally pooled models result in correctly specified densities according to all tests.

[19] The KS test only detects mis-specification in the BMA-MC3 and the BVAR.

Table 3 shows results for medium horizon (one-year-ahead) predictive densities. Due to the maintained assumption of independence and the serial correlation built, by construction, in the four-quarter-ahead forecasts, we divide the out-of-sample period into four subsets whose observations are 4 periods apart. For brevity, we report the minimum p-value across the various subsets. The left panel shows that only ADL models using exchange rates as predictors result in mis-specified densities and only according to the AD test. Furthermore, both the KS and AD tests find empirical evidence against the correct specification of the BMA-MC3 and BVAR models. The right panel shows that there is more evidence of mis-specification for the ADL models when forecasting inflation rather than output growth at medium horizons: several nominal interest rate measures do not result in correctly specified densities. The tests reject the correct specification of several forecast combination models (including BMS-OLS, BMA-MC3 and the BVAR models), while do not reject correct specification for the simple average and BMA models.

INSERT TABLES 2 AND 3 HERE

Overall, by comparing the right and left panels in the tables, under the maintained assumption of independence, there is more empirical evidence against correct specification for density forecasts of inflation than for output growth, at both short and medium horizons. By comparing ADL models across Tables 2 and 3, we conclude that normality is more appropriate for forecasting one-year-ahead than one-quarter-ahead output growth and inflation. Regarding model combinations, the most robust result is that normality cannot be rejected for the simple average and BMA models across horizons (with the exception of BMA for forecasting output growth at short horizons). The factor model also performs well in all cases but forecasting inflation at the one-quarter-ahead horizon.

4.2 Tests of Independence

Correct specification of density forecasts also requires independence of the PITs. Tables 4 and 5 report results for the Ljung-Box (LB) test of no-autocorrelation in the PITs. Table 4 focuses on forecast horizons equal to one quarter ($h = 1$). The left panel in Table 4 reports results for forecasting output growth and the right panel reports results for forecasting inflation. For each of the models, reported in the first column of each panel, the tables report the p-values of the LB test for serial correlation in the mean (next column) and in the variance of $\{z_{t+h}\}_{t=R}^{T}$ (second to next column).

For output growth, Table 4 shows very little statistical evidence of serial correlation in the first moments of the PITs (except for the BMA-MC3 model and for the ADL models with the T-bill rate). There is significant serial correlation in the second moments of the PITs for the ADL model for several predictors (especially medium and long interest rates and some measures of money), as well as for the BMA-MC3 and BVAR models. The equal average, BMA-OLS, BMA and factor models show no serial correlation in either the first or the second moments of their PITs.

Turning to inflation, reported on the right panel of Table 4, the striking result is that serial correlation in the second moments of the PITs is rejected for most of the ADL models (with the exception of real overnight interest rates, earnings and real M3 measures), as well as for most density combinations (with the exception of BMA-OLS and BMA). Instead, there is no evidence of serial correlation in the first moments of the PITs for most ADL models, nor for the simple average and the factor models; however, there is serial correlation in the PITs of BMA-MC3 and BVAR models.

INSERT TABLES 4 AND 5 HERE

Table 5 reports results for one-year-ahead density forecasts ($h = 4$). Due to the serial correlation built by construction in the four-step-ahead forecasts, we divide the out-of-sample period into four subsets whose observations are 4 periods apart. For brevity, we report the minimum p-value across the various subsets. Table 5 shows very little evidence of serial correlation in the PITs for output growth across various specification, and reports a few rejections of no serial correlation in the PITs of inflation. For output growth, almost all of the ADL, simple average, BMA-OLS, BMA models, factor, as well as BVAR models are correctly specified; however, the test rejects independence in the PITs of the BMA-MC3 model. There is slightly stronger evidence of serial correlation in PITs of inflation forecasts, (especially in the second moments of the PITs) for the ADL models with employment, unemployment, as well as several money and interest rate measures. The simple average, BMA-MC3, factor and BVAR models also result in mis-calibrated densities. Instead, the BMA and BMA-OLS models do not show evidence of serial correlation in the PITs.

In general, forecast densities are fairly well calibrated in terms of uncorrelation in the PITs for GDP growth, with occasional exceptions for the ADL model with selected predictors, but less so for inflation. In addition, there is more evidence of correlation in the PITs for one-quarter-ahead forecast densities than in one-year-ahead ones, as well second moments versus the first. The most robust result in favor of correct specification across

horizons and predictors refers, again, to equal averaging and BMA models, although simple averaging shows evidence of serial correlation in the second moments of the PITs in the case of inflation.[20]

4.3 Tests of Identical Distribution

There is empirical evidence in the forecasting literature that predictors' Granger-causality is unstable over time: see Stock and Watson (1996, 2003, 2007) and Rossi (2013). Here, we are concerned that the distribution of the PITs might have changed over time. We investigate the stability of the first and second (non-central) moments of the PITs using Andrews' (1993) test. Tables 6 and 7 provide the results for one- and four-quarter-ahead forecast horizons, respectively, where, again, for the case of $h = 4$, we report the minimum p-value across the various independent subsets $h-1$ periods apart. Table 6 shows that we reject the stability of the PITs of output growth for a few nominal interest rate predictors in the ADL model, as well as in the BMA-MC3 model. There is less evidence of instabilities in density forecasts of inflation. As Table 7 suggests, there is stronger evidence of instabilities in the four-quarter-ahead predictive densities of both output growth and inflation: the test detects instabilities for the ADL model when several predictors are used (e.g. interest rates and real money when predicting output growth and exchange rates, capacity utilization and M1 when predicting inflation). The instabilities mostly affect the second (non-central) moments of the PITs. In addition, there is evidence of instability in the predictive density of BMA, BMA-MC3, as well as BVAR models when predicting output growth. On the other hand, there is no evidence of instability in predictive densities of simple average, BMA-OLS and factor models. For the case of inflation, models based on pooling result in stable densities as well, with the exception of BMA-OLS and BMA-MC3 models. The break dates reported for $h = 1$ correspond to the Great Moderation. Given that the sub-sample based analysis relies on Bonferroni bounds, we do not report break dates for $h = 4$.

INSERT TABLES 6 AND 7 HERE

[20] As an alternative, one could implement the BDS test by Broock, Scheinkman and Dechert (1987). The BDS test is a non-parametric test of the null hypothesis of independent and identical distribution against an unspecified alternative and operates by reshuffling the observations.

4.4 Tests on the Inverse Normal of the PIT

Finally, we report results for tests based on the inverse normal of the PIT. Recall that, according to Berkowitz (2001), the latter not only can test jointly for uniformity and serial correlation, but are also more powerful than the previous ones we reported. Tables 8 and 9 report results for Berkowitz's (2001) tests whereas Table 10 reports results for the Doornik and Hansen (2008) test.

Interestingly, at all horizons, Tables 8 and 9 show that there is strong evidence of mis-specification in the PITs for both output growth and inflation according to Berkowitz's (2001) test for uniformity (labeled "$\mu = 0$, $\sigma = 1$"). Basically, the only models that are not mis-specified for forecasting output growth at short horizons are the ADL model with exchange rate measures as well as the simple average, and none of the models for predicting inflation at short horizons, except BMA. On the other hand, at the one-year-ahead horizon, the ADL models are all correctly specified for output growth, as well as the BMA, BMA-OLS, factor and simple average models. When predicting inflation one-year-ahead, only ADL models based on nominal interest rates, as well as the BMA-OLS, BMA-MC3, and BVAR models are not correctly specified.

INSERT TABLES 8 AND 9 HERE

We should note that Tables 8 and 9 also provide evidence on lack of serial correlation in the PITs (columns 3 and 7, labeled "$\rho = 0$"), as well as against the joint hypothesis of independence and normality of inverse normal transform of the PITs (columns 4 and 8, labeled "joint"). The results of no serial correlation are in-line with the ones implied by the Ljung-Box test reported in Tables 4 and 5. Serial correlation in the first moment of the PITs is almost inexistent for both short and medium horizon predictive densities for both output growth and inflation. The joint hypothesis is rejected for several models for both inflation and output growth, primarily at the one-quarter-ahead forecast horizon. By comparing columns two and four (and six with eight), it appears that the joint hypothesis results mostly imitate those of the test of uniformity.

Finally, Doornik and Hansen's (2008) test, which relies on transformed skewness and kurtosis measures, does not detect strong mis-specification in the predictive densities of several ADL, BMA-OLS and factor models. However, based on this test, the simple average, BMA, BMA-MC3 as well as BVAR models appear to be mis-specified. Notably, the evidence of improper calibration is stronger for one-step-ahead density forecasts relative to the one-year-ahead ones.

INSERT TABLE 10 HERE

4.5 A Summary of the Empirical Results

Table 11 provides a summary of the empirical results across models and test statistics. For each model and for each test, it summarizes the empirical evidence on the property listed in the corresponding column. For example, for the tests for uniformity listed in columns 2-5, "yes" denotes that uniformity is not rejected at the 5% significance level. The table shows that, for many models, the assumption of normality of density forecasts is mis-specified, according to at least one of the tests we consider. The evidence in favor of correct specification of normality is the strongest for equally-weighted forecast averages, especially for predicting output growth at ones-year-ahead horizon, in which case none of the tests rejects correct specification. The same holds for BMA for one-quarter-ahead inflation density forecasts. Overall, the performance of both models is more robust across target variables and horizons than that of all the other models we consider. Thus, while each of the ADL models is mis-specified for some predictors and according to some tests, their average is not. This suggests that non-normality is important, except possibly for equal average and BMA density forecasts combination models.

INSERT TABLE 11 HERE

5 Conclusions

This paper evaluates the correct specification of predictive densities of U.S. inflation and output growth, based on an extensive data set of macroeconomic predictors. Our empirical findings show that, according to most tests, predictive densities of predictive density combinations based on simple, equal weighting, as well as Bayesian Model Averaging appear to be one of the best calibrated models in terms of normality. We conjecture that averaging across series and models might be the reason for this result. Whether or not normality is an appropriate assumption for each individual ADL model crucially depends on the predictor, although most predictors typically fail according to at least one of the tests. The results for the factor and BVAR model-based, as well as the alternative ways of combining densities considered in this paper are much less robust: the normality assumption is rejected according to several tests, at least at some forecast horizons.

References

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

[2] Anderson, T.W. and D.A. Darling (1952), "Asymptotic Theory of Certain "Goodness-of-Fit" Criteria based on Stochastic Processes", *Annals of Mathematical Statistics* 23(2), 193–212.

[3] Anderson, T.W. and D.A. Darling (1954), "A Test of Goodness-of-Fit", *Journal of the American Statistical Association* 49(268), 765–769.

[4] Andrews, D.W.K. (1993), "Tests for Parameter Instability and Structural Change with Unknown Change Point", *Econometrica* 61(4), 821–856.

[5] Bai, J. and S. Ng (2002), "Determining the Number of Factors in Approximate Factor Models", *Econometrica* 70(1), 191-221.

[6] Bai, J. and S. Ng (2007), "Determining the Number of Primitive Shocks in Factor Models", *Journal of Business and Economic Statistics* 25(1), 52-60.

[7] Banerjee, A. and M. Marcellino (2006) "Are There any Reliable Leading Indicators for the US Inflation and GDP Growth?", *International Journal of Forecasting* 22, 137-151.

[8] Banerjee, A., M. Marcellino and I. Masten (2005), "Leading Indicators for Euro-Area Inflation and GDP Growth", *Oxford Bulletin of Economics and Statistics* 67(S1), 785-813.

[9] Berkowitz, J. (2001), "Testing Density Forecasts, with Applications to Risk Management", *Journal of Business and Economic Statistics* 19(4), 465-474.

[10] Broock, W. A., J.A. Scheinkman and W.D. Dechert (1996), "A Test for Independence based on the Correlation Dimension", *Econometric Reviews* 15(3), 197-235.

[11] Clark, T.E. (2011), "Real-Time Density Forecasts From Bayesian Vector Autoregressions With Stochastic Volatility", *Journal of Business and Economic Statistics* 29(3), 327-341.

[12] Clements, M.P. and J. Smith (2000), "Evaluating the Forecast Densities of Linear and Non-linear Models: Applications to Output Growth and Unemployment", *Journal of Forecasting* 19(4), 255-276.

[13] Corradi, V. and N.R. Swanson (2006b), "Predictive Density Evaluation", in: G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Vol. 1, North Holland: Elsevier, 197-284.

[14] Diks, C., V. Panchenko and D. van Dijk (2011), "Likelihood-based Scoring Rules for Comparing Density Forecasts in Tails", *Journal of Econometrics* 163(2), 215–230.

[15] Diebold, F.X., T.A. Gunther, and A.S. Tay (1998), "Evaluating Density Forecasts with Applications to Financial Risk Management", *International Economic Review* 39(4), 863-883.

[16] Diebold F.X., A.S. Tay and K.F. Wallis (1999), "Evaluating Density Forecasts of Inflation: the Survey of Professional Forecasters", in: R.F. Engle and H. White, *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger*, Oxford University Press, 76-90.

[17] Doornik, J.A. and H. Hansen (2008), "An Omnibus Test for Univariate and Multivariate Normality", *Oxford Bulletin of Economics and Statistics* 70(S1), 927-939.

[18] Garratt, A., K. Lee, M.H. Pesaran and Y. Shin (2003), "Forecast Uncertainties in Macroeconomic Modeling: An Application to the U.K. Economy", *Journal of the American Statistical Association* 98(464), 829-838.

[19] Giannone, D., M. Lenza and G. Primiceri (2012), "Prior Selection for Vector Autoregressions ", *mimeo*.

[20] Granger, C. and M.H. Pesaran (2000), "Economic and Statistical Measures of Forecast Accuracy", *Journal of Forecasting* 19(7), 537-560.

[21] Guidolin, M. and A. Timmermann (2006), "Term Structure of Risk Under Alternative Econometric Specifications", *Journal of Econometrics* 131(1-2), 285-308.

[22] Hayashi, F. (2000), *Econometrics*, Princeton University Press.

[23] Jore, A.S., J. Mitchell and S.P. Vahey (2010), "Combining Forecast Densities From VARs with Uncertain Instabilities", *Journal of Applied Econometrics* 25(4), 621-634.

[24] Kolmogorov, A.N. (1933), "Sulla Determinazione Empirica di una Legge di Distribuzione", *Giornale dell'Istituto Italiano Degli Attuari* 4, 83-91.

[25] Kroese, D.P., T. Taimre and Z.I. Botev (2011), *Handbook of Monte Carlo Methods*, John Wiley.

[26] Koop, G. (2003), *Bayesian Econometrics*, John Wiley.

[27] Manzan, S. and D. Zerom (2013), "Are Macroeconomic Variables Useful for Forecasting the Distribution of U.S. Inflation?", *International Journal of Forecasting*, forthcoming.

[28] Marcellino, M., Stock J.H. and M.W. Watson (2003), "Macroeconomic Forecasting in the Euro area: Country Specific Versus Euro Wide Information", *European Economic Review* 47(1), 1-18.

[29] Mitchell, J. and K.F. Wallis. (2011), "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness", *Journal of Applied Econometrics* 26(6), 1023-1040.

[30] Newey, W.K. and K.O. West (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica* 55(3), 703-708.

[31] Noceti, P., J. Smith and S. Hodges, (2003), "An Evaluation of Tests of Distributional Forecasts", *Journal of Forecasting* 22(6-7), 447-455.

[32] Rosenblatt, M. (1952), "Remarks on a Multivariate Transformation", *Annals of Mathematical Statistics* 23(3), 470-472.

[33] Rossi, B. (2013), "Advances in Forecasting Under Instabilities", in G. Elliott and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Volume 2, Elsevier-North Holland Publications.

[34] Rossi, B. and T. Sekhposyan (2010), "Have Economic Models' Forecasting Performance for US Output Growth and Inflation Changed Over Time, and When?", *International Journal of Forecasting* 26(4), 808-835.

[35] Rossi, B. and T. Sekhposyan (2012), "Conditional Predictive Density Evaluation in the Presence of Instabilities", *Journal of Econometrics,* forthcoming.

[36] Sims, C. A. and T. Zha (1998), "Bayesian Methods for Dynamic Multivariate Models", *International Economic Review* 39(4), 949-968.

[37] Smirnov, N.V. (1948), "Tables for Estimating the Goodness of Fit of Empirical Distributions", *The Annals of Mathematical Statistics* 19(2), 279-281.

[38] Stock, J.H. and M.W. Watson (1996), "Evidence on Structural Instability in Macroeconomic Time Series Relations", *Journal of Business and Economic Statistics* 14(1), 11-30.

[39] Stock, J.H. and M.W. Watson (2002), *Introduction to Econometrics,* Addison-Wesley.

[40] Stock, J.H. and M.W. Watson (2003), "Forecasting Output and Inflation: The Role of Asset Prices", *Journal of Economic Literature* 41(3)*, 788-829.*

[41] Stock, J.H. and M.W. Watson (2004), "Combination Forecasts of Output Growth in a Seven Country Data Set", *Journal of Forecasting* 23(6), 405-430.

[42] Stock, J.H. and M.W. Watson (2007), "Has Inflation Become Harder to Forecast?", *Journal of Money, Credit and Banking* 39(1), 3-33.

[43] Timmermann, A. (2006), "Forecast Combinations", in: G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting* Vol. 1, North Holland: Elsevier, 135-196.

[44] Wright, J.H. (2009), "Forecasting US Inflation by Bayesian Model Averaging", *Journal of Forecasting* 28(2), 131-144.

6 Figures and Tables

Table 1. Description of Data Series

Label	Trans	Period	Name	Description	Source
			Asset Prices		
rovnght@us	level	59:M1-10:M9	FEDFUNDS	Int. Rate: Fed Funds (Effective)	F
rtbill@us	level	59:M1-10:M9	TB3MS	Int. Rate: 3-Mn Tr. Bill, Sec Mkt Rate	F
rbnds@us	level	59:M1-10:M9	GS1	Int. Rate: US Tr. Const Mat., 1-Yr	F
rbndm@us	level	59:M1-10:M9	GS5	Int. Rate: US Tr. Const Mat., 5-Yr	F
rbndl@us	level	59:M1-10:M9	GS10	Int. Rate: US Tr. Const Mat., 10-Yr	F
stockp@us	Δln	59:Q1-10:Q3	SP500	US Share Prices: S&P 500	F
exrate@us	Δln	73:M1-10:M9	111..NELZF...	NEER from UCL	I
			Real Activity		
rgdp@us	Δln	59:Q1-10:Q3	GDPC96	Real GDP, sa	F
ip@us	Δln	59:M1-10:M9	INDPRO	Industrial Production Index, sa	F
capu@us	level	59:M1-10:M9	CAPUB04	Capacity Utilization Rate: Man., sa	F
emp@us	Δln	59:M1-10:M9	CE16OV	Civilian Employment: thsnds, sa	F
unemp@us	level	59:M1-10:M9	UNRATE	Civilian Unemployment Rate, sa	F
			Wages and prices		
pgdp@us	Δln	59:Q1-10:Q3	GDPDEF	GDP Deflator, sa	F
cpi@us	Δln	59:M1-10:M9	CPIAUCSL	CPI: Urban, All items, sa	F
ppi@us	Δln	59:M1-10:M9	PPIACO	Producer Price Index, nsa	F
earn@us	Δln	59:M1-10:M9	AHEMAN	Hourly Earnings: Man., nsa	F
			Money		
mon0@us	Δln	59:M1-10:M9	AMBSL	Monetary Base, sa	F
mon1@us	Δln	59:M1-10:M9	M1SL	Money: M1, sa	F
mon2@us	Δln	59:M1-10:M9	M2SL	Money: M2, sa	F
mon3@us	Δln	59:M1-06:M2	M3SL	Money: M3, sa	F

Notes: Sources are abbreviated as follows: “F" - Federal Reserve Economic Data (FRED) and “I" - IMF International Financial Statistics. When the names in the table are preceded with a prefix “r", it indicates real variable adjusted either by the CPI (stock variables) or CPI inflation (flow variables). Interest rate spread is calculated as the difference between “rbndl" and “rovnght".

Table 2. Tests of Correct Specification at $h = 1$

Output Growth			Inflation		
Variable	KS	AD	Variable	KS	AD
rgdp@us	0.30	0.41	pgdp@us	0.48	0.03 *
rovnght@us	0.33	0.06	rovnght@us	0.27	0.00 *
rtbill@us	0.46	0.09	rtbill@us	0.24	0.00 *
rbnds@us	0.51	0.06	rbnds@us	0.29	0.00 *
rbndm@us	0.79	0.09	rbndm@us	0.34	0.01 *
rbndl@us	0.70	0.42	rbndl@us	0.38	0.01 *
rspread@us	0.21	0.02 *	rspread@us	0.23	0.01 *
stockp@us	0.24	0.02 *	stockp@us	0.49	0.01 *
exrate@us	0.57	0.00 *	exrate@us	0.78	0.00 *
rrovnght@us	0.33	0.06	rrovnght@us	0.29	0.01 *
rrtbill@us	0.50	0.44	rrtbill@us	0.44	0.01 *
rrbnds@us	0.35	0.43	rrbnds@us	0.47	0.01 *
rrbndm@cn	0.39	0.03 *	rrbndm@cn	0.33	0.01 *
rrbndl@us	0.40	0.02 *	rrbndl@us	0.25	0.00 *
rstockp@us	0.08	0.01 *	rstockp@us	0.47	0.01 *
rexrate@us	0.57	0.00 *	rexrate@us	0.78	0.00 *
ip@us	0.29	0.42	rgdp@us	0.28	0.01 *
capu@us	0.23	0.04 *	ip@us	0.53	0.01 *
emp@us	0.33	0.37	capu@us	0.22	0.01 *
unemp@us	0.03 *	0.01 *	emp@us	0.27	0.01 *
pgdp@us	0.39	0.03 *	unemp@us	0.18	0.01 *
cpi@us	0.25	0.03 *	cpi@us	0.27	0.00 *
ppi@us	0.27	0.03 *	ppi@us	0.34	0.01 *
earn@us	0.28	0.01 *	earn@us	0.16	0.01 *
mon0@us	0.41	0.08	mon0@us	0.28	0.00 *
mon1@us	0.22	0.02 *	mon1@us	0.35	0.01 *
mon2@us	0.32	0.03 *	mon2@us	0.22	0.00 *
mon3@us	0.27	0.00 *	mon3@us	0.33	0.01 *
rmon0@us	0.27	0.01 *	rmon0@us	0.33	0.00 *
rmon1@us	0.27	0.05	rmon1@us	0.30	0.00 *
rmon2@us	0.59	0.08	rmon2@us	0.30	0.01 *
rmon3@us	0.65	0.02 *	rmon3@us	0.74	0.03 *
Simple Average	0.37	0.35	Simple Average	0.56	0.05
BMA-OLS	0.14	0.01 *	BMA-OLS	0.19	0.00 *
BMA	0.11	0.03 *	BMA	0.79	0.39
BMA-MC3	0.00 *	0.00 *	BMA-MC3	0.00 *	0.00 *
Factor	0.40	0.05	Factor	0.10	0.00 *
BVAR	0.00 *	0.00 *	BVAR	0.01 *	0.00 *

Notes: We approximate the critical values of KS and AD tests as in Kroese et al (2011).

* marks rejection at a 5% significance level.

Table 3. Tests of Correct Specification at $h = 4$

Output Growth			Inflation		
Variable	KS	AD	Variable	KS	AD
rgdp@us	0.24	0.35	pgdp@us	0.36	0.36
rovnght@us	0.43	0.38	rovnght@us	0.04	0.00 *
rtbill@us	0.49	0.39	rtbill@us	0.01 *	0.00 *
rbnds@us	0.34	0.35	rbnds@us	0.02	0.00 *
rbndm@us	0.73	0.41	rbndm@us	0.04	0.00 *
rbndl@us	0.78	0.53	rbndl@us	0.07	0.01 *
rspread@us	0.28	0.35	rspread@us	0.35	0.05
stockp@us	0.36	0.39	stockp@us	0.45	0.37
exrate@us	0.46	0.00 *	exrate@us	0.25	0.00 *
rrovnght@us	0.36	0.35	rrovnght@us	0.55	0.36
rrtbill@us	0.55	0.40	rrtbill@us	0.40	0.38
rrbnds@us	0.47	0.37	rrbnds@us	0.47	0.38
rrbndm@cn	0.48	0.36	rrbndm@cn	0.41	0.36
rrbndl@us	0.49	0.35	rrbndl@us	0.37	0.09
rstockp@us	0.34	0.39	rstockp@us	0.46	0.37
rexrate@us	0.46	0.00 *	rexrate@us	0.25	0.00 *
ip@us	0.28	0.36	rgdp@us	0.52	0.36
capu@us	0.42	0.36	ip@us	0.30	0.36
emp@us	0.27	0.35	capu@us	0.29	0.37
unemp@us	0.19	0.08	emp@us	0.13	0.08
pgdp@us	0.56	0.38	unemp@us	0.32	0.36
cpi@us	0.58	0.35	cpi@us	0.28	0.05
ppi@us	0.64	0.37	ppi@us	0.38	0.35
earn@us	0.26	0.36	earn@us	0.06	0.03
mon0@us	0.49	0.36	mon0@us	0.33	0.36
mon1@us	0.31	0.36	mon1@us	0.21	0.02
mon2@us	0.42	0.35	mon2@us	0.24	0.08
mon3@us	0.25	0.04	mon3@us	0.08	0.02
rmon0@us	0.23	0.06	rmon0@us	0.49	0.35
rmon1@us	0.59	0.37	rmon1@us	0.12	0.35
rmon2@us	0.68	0.41	rmon2@us	0.18	0.36
rmon3@us	0.62	0.08	rmon3@us	0.76	0.08
Simple Average	0.15	0.36	Simple Average	0.34	0.36
BMA-OLS	0.22	0.37	BMA-OLS	0.01 *	0.00 *
BMA	0.43	0.35	BMA	0.25	0.35
BMA-MC3	0.00 *	0.00 *	BMA-MC3	0.05	0.01 *
Factor	0.51	0.38	Factor	0.42	0.04
BVAR	0.00 *	0.00 *	BVAR	0.02	0.00 *

Notes: The table reports minimum p-values of the KS and AD tests (approximated as in Kroese et al (2011)) based on four subsets $\{z_1, z_{1+h}, z_{1+2h}, ...\}$,$\{z_2, z_{2+h}, z_{2+2h}, ...\}$,$\{z_3, z_{3+h}, z_{3+2h}, ...\}$, and $\{z_4, z_{4+h}, z_{4+2h}, ...\}$.

* indicates rejection at 5% significance with Bonferroni bounds.

Table 4. Ljung-Box test at $h = 1$

Output Growth			Inflation		
Variable	$(z_{t+h} - \bar{z})$	$(z_{t+h} - \bar{z})^2$	Variable	$(z_{t+h} - \bar{z})$	$(z_{t+h} - \bar{z})^2$
rgdp@us	0.41	0.01 *	pgdp@us	0.52	0.01 *
rovnght@us	0.18	0.35	rovnght@us	0.33	0.01 *
rtbill@us	0.04 *	0.12	rtbill@us	0.36	0.03 *
rbnds@us	0.19	0.16	rbnds@us	0.44	0.00 *
rbndm@us	0.14	0.03 *	rbndm@us	0.37	0.01 *
rbndl@us	0.21	0.02 *	rbndl@us	0.31	0.01 *
rspread@us	0.27	0.16	rspread@us	0.25	0.01 *
stockp@us	0.87	0.26	stockp@us	0.67	0.00 *
exrate@us	0.48	0.31	exrate@us	0.37	0.01 *
rrovnght@us	0.41	0.00 *	rrovnght@us	0.56	0.09
rrtbill@us	0.28	0.04 *	rrtbill@us	0.41	0.00 *
rrbnds@us	0.30	0.05	rrbnds@us	0.38	0.00 *
rrbndm@cn	0.38	0.01 *	rrbndm@cn	0.33	0.00 *
rrbndl@us	0.35	0.00 *	rrbndl@us	0.44	0.00 *
rstockp@us	0.84	0.17	rstockp@us	0.66	0.00 *
rexrate@us	0.48	0.31	rexrate@us	0.37	0.01 *
ip@us	0.23	0.42	rgdp@us	0.27	0.00 *
capu@us	0.45	0.32	ip@us	0.55	0.01 *
emp@us	0.31	0.12	capu@us	0.28	0.01 *
unemp@us	0.55	0.06	emp@us	0.36	0.00 *
pgdp@us	0.07	0.07	unemp@us	0.47	0.01 *
cpi@us	0.43	0.12	cpi@us	0.14	0.00 *
ppi@us	0.94	0.14	ppi@us	0.19	0.00 *
earn@us	0.50	0.07	earn@us	0.47	0.07
mon0@us	0.69	0.03 *	mon0@us	0.79	0.02 *
mon1@us	0.05	0.06	mon1@us	0.60	0.02 *
mon2@us	0.76	0.00 *	mon2@us	0.73	0.02 *
mon3@us	0.49	0.00 *	mon3@us	0.40	0.36
rmon0@us	0.68	0.09	rmon0@us	0.88	0.00 *
rmon1@us	0.08	0.55	rmon1@us	0.51	0.00 *
rmon2@us	0.69	0.00 *	rmon2@us	0.77	0.00 *
rmon3@us	0.55	0.00 *	rmon3@us	0.44	0.17
Simple Average	0.63	0.10	Simple Average	0.30	0.03 *
BMA-OLS	0.59	0.31	BMA-OLS	0.12	0.08
BMA	0.15	0.14	BMA	0.06	0.07
BMA-MC3	0.00 *	0.00 *	BMA-MC3	0.00 *	0.00 *
Factor	0.44	0.66	Factor	0.24	0.00 *
BVAR	0.20	0.00 *	BVAR	0.00 *	0.00 *

Notes: The table reports p-values of the LB test based on a $\chi^2(4)$. * indicates rejection at 5% significance level.

Table 5. Ljung-Box test at $h = 4$

Output Growth			Inflation		
Variable	$(z_{t+h} - \bar{z})$	$(z_{t+h} - \bar{z})^2$	Variable	$(z_{t+h} - \bar{z})$	$(z_{t+h} - \bar{z})^2$
rgdp@us	0.10	0.19	pgdp@us	0.11	0.00 *
rovnght@us	0.01 *	0.08	rovnght@us	0.17	0.03
rtbill@us	0.00 *	0.02	rtbill@us	0.21	0.11
rbnds@us	0.05	0.02	rbnds@us	0.18	0.00 *
rbndm@us	0.24	0.01 *	rbndm@us	0.14	0.04
rbndl@us	0.22	0.04	rbndl@us	0.09	0.01 *
rspread@us	0.53	0.38	rspread@us	0.09	0.06
stockp@us	0.25	0.34	stockp@us	0.21	0.03
exrate@us	0.54	0.14	exrate@us	0.59	0.07
rrovnght@us	0.12	0.11	rrovnght@us	0.32	0.06
rrtbill@us	0.21	0.09	rrtbill@us	0.15	0.12
rrbnds@us	0.28	0.45	rrbnds@us	0.41	0.14
rrbndm@cn	0.56	0.06	rrbndm@cn	0.27	0.02
rrbndl@us	0.46	0.03	rrbndl@us	0.29	0.14
rstockp@us	0.28	0.43	rstockp@us	0.21	0.03
rexrate@us	0.54	0.14	rexrate@us	0.59	0.07
ip@us	0.19	0.64	rgdp@us	0.05	0.03
capu@us	0.08	0.15	ip@us	0.12	0.03
emp@us	0.11	0.12	capu@us	0.27	0.08
unemp@us	0.23	0.45	emp@us	0.16	0.01 *
pgdp@us	0.57	0.26	unemp@us	0.05	0.00 *
cpi@us	0.19	0.16	cpi@us	0.31	0.07
ppi@us	0.29	0.19	ppi@us	0.13	0.00 *
earn@us	0.29	0.43	earn@us	0.07	0.00 *
mon0@us	0.26	0.14	mon0@us	0.07	0.00 *
mon1@us	0.16	0.34	mon1@us	0.07	0.00 *
mon2@us	0.57	0.12	mon2@us	0.13	0.02
mon3@us	0.20	0.02	mon3@us	0.02	0.22
rmon0@us	0.50	0.47	rmon0@us	0.26	0.00 *
rmon1@us	0.16	0.64	rmon1@us	0.02	0.01 *
rmon2@us	0.37	0.10	rmon2@us	0.05	0.01 *
rmon3@us	0.32	0.07	rmon3@us	0.04	0.04
Simple Average	0.46	0.36	Simple Average	0.16	0.01 *
BMA-OLS	0.35	0.49	BMA-OLS	0.29	0.28
BMA	0.34	0.11	BMA	0.15	0.12
BMA-MC3	0.00 *	0.13	BMA-MC3	0.00 *	0.01 *
Factor	0.70	0.05	Factor	0.38	0.00 *
BVAR	0.03	0.66	BVAR	0.03	0.00 *

Notes: The table reports minimum p-values of the LB test based on a $\chi^2(4)$ for four subsets $\{z_1, z_{1+h}, z_{1+2h}, ...\}$, $\{z_2, z_{2+h}, z_{2+2h}, ...\}$, $\{z_3, z_{3+h}, z_{3+2h}, ...\}$, and $\{z_4, z_{4+h}, z_{4+2h}, ...\}$.

* indicates rejection at 5% significance with Bonferroni bounds.

Table 6. Andrews (1993) QLR test at $h = 1$

	Output Growth					Inflation			
Variable	z_{t+h}		z^2_{t+h}		Variable	z_{t+h}		z^2_{t+h}	
rgdp@us	0.78		0.89		pgdp@us	1.00		1.00	
rovnght@us	0.02 *	1985:III	0.01 *	1985:I	rovnght@us	1.00		0.86	
rtbill@us	0.06		0.04 *	1985:III	rtbill@us	1.00		0.80	
rbnds@us	0.12		0.07		rbnds@us	1.00		0.70	
rbndm@us	0.06		0.05		rbndm@us	1.00		0.81	
rbndl@us	0.16		0.11		rbndl@us	1.00		0.75	
rspread@us	1.00		1.00		rspread@us	1.00		1.00	
stockp@us	1.00		1.00		stockp@us	1.00		1.00	
exrate@us	0.39		0.67		exrate@us	1.00		1.00	
rrovnght@us	0.57		0.64		rrovnght@us	1.00		1.00	
rrtbill@us	0.60		0.79		rrtbill@us	1.00		1.00	
rrbnds@us	0.69		0.84		rrbnds@us	1.00		1.00	
rrbndm@cn	0.66		0.80		rrbndm@cn	1.00		1.00	
rrbndl@us	0.78		0.89		rrbndl@us	1.00		1.00	
rstockp@us	1.00		1.00		rstockp@us	1.00		1.00	
rexrate@us	0.39		0.67		rexrate@us	1.00		1.00	
ip@us	0.49		0.63		rgdp@us	1.00		0.84	
capu@us	0.46		0.35		ip@us	1.00		0.79	
emp@us	0.89		1.00		capu@us	1.00		0.63	
unemp@us	0.29		0.56		emp@us	1.00		0.82	
pgdp@us	0.11		0.07		unemp@us	0.57		0.43	
cpi@us	0.19		0.12		cpi@us	1.00		1.00	
ppi@us	1.00		1.00		ppi@us	0.76		0.69	
earn@us	0.87		0.87		earn@us	1.00		1.00	
mon0@us	0.66		0.68		mon0@us	1.00		0.85	
mon1@us	0.50		0.66		mon1@us	1.00		1.00	
mon2@us	0.60		0.85		mon2@us	1.00		0.58	
mon3@us	0.88		0.85		mon3@us	1.00		0.68	
rmon0@us	0.89		1.00		rmon0@us	1.00		0.85	
rmon1@us	0.89		0.81		rmon1@us	1.00		0.57	
rmon2@us	0.67		0.84		rmon2@us	1.00		0.59	
rmon3@us	0.66		0.61		rmon3@us	1.00		0.66	
Simple Average	0.65		0.82		Simple Average	1.00		1.00	
BMA-OLS	0.84		0.56		BMA-OLS	1.00		1.00	
BMA	0.60		0.28		BMA	1.00		1.00	
BMA-MC3	0.00 *	1988:III	0.00 *	1988:III	BMA-MC3	0.00 *	1975:IV	0.00 *	1975:IV
Factor	0.87		0.74		Factor	0.85		0.38	
BVAR	0.37		0.63		BVAR	1.00		0.85	

Notes: The table reports p-values and break dates of the Andrews QLR test. * indicates rejection at 5% significance level.

Table 7. Andrews (1993) QLR test at $h = 4$

Output Growth			Inflation		
Variable	z_{t+h}	z^2_{t+h}	Variable	z_{t+h}	z^2_{t+h}
rgdp@us	0.36	0.19	pgdp@us	0.52	1.00
rovnght@us	0.00 *	0.00 *	rovnght@us	0.44	0.59
rtbill@us	0.00 *	0.00 *	rtbill@us	0.76	1.00
rbnds@us	0.00 *	0.00 *	rbnds@us	0.72	0.89
rbndm@us	0.02	0.00 *	rbndm@us	0.04	0.03
rbndl@us	0.01 *	0.00 *	rbndl@us	0.19	0.16
rspread@us	0.51	0.23	rspread@us	0.08	0.44
stockp@us	0.18	0.02	stockp@us	0.49	0.24
exrate@us	0.06	0.05	exrate@us	0.00 *	0.00 *
rrovnght@us	0.32	0.09	rrovnght@us	0.54	0.35
rrtbill@us	0.16	0.00 *	rrtbill@us	0.73	0.46
rrbnds@us	0.11	0.00 *	rrbnds@us	0.64	0.71
rrbndm@cn	0.06	0.00 *	rrbndm@cn	0.63	0.89
rrbndl@us	0.06	0.00 *	rrbndl@us	0.06	0.02
rstockp@us	0.16	0.02	rstockp@us	0.46	0.21
rexrate@us	0.06	0.05	rexrate@us	0.00 *	0.00 *
ip@us	0.65	0.24	rgdp@us	0.40	0.70
capu@us	0.10	0.00 *	ip@us	0.54	0.68
emp@us	0.44	0.30	capu@us	0.01 *	0.00 *
unemp@us	0.05	0.00 *	emp@us	0.27	0.60
pgdp@us	0.08	0.14	unemp@us	0.73	0.51
cpi@us	0.00 *	0.00 *	cpi@us	0.33	0.09
ppi@us	0.20	0.01 *	ppi@us	0.49	0.83
earn@us	0.69	0.34	earn@us	0.56	1.00
mon0@us	0.06	0.02	mon0@us	0.22	0.64
mon1@us	0.39	0.18	mon1@us	0.00 *	0.00 *
mon2@us	0.29	0.04	mon2@us	0.52	0.82
mon3@us	0.54	0.06	mon3@us	0.81	0.81
rmon0@us	0.25	0.08	rmon0@us	0.14	0.05
rmon1@us	0.20	0.00 *	rmon1@us	0.00 *	0.00 *
rmon2@us	0.05	0.00 *	rmon2@us	0.33	0.43
rmon3@us	0.21	0.00 *	rmon3@us	0.45	0.88
Simple Average	0.22	0.08	Simple Average	0.38	0.23
BMA-OLS	0.07	0.03	BMA-OLS	0.02	0.00 *
BMA	0.02	0.00 *	BMA	0.51	0.35
BMA-MC3	0.00 *	0.00 *	BMA-MC3	0.00 *	0.00 *
Factor	0.18	0.34	Factor	0.76	0.45
BVAR	0.00 *	0.00 *	BVAR	0.05	0.02

Notes: The table reports minimum p-values of the Andrews QLR test based on four subsets $\{z_1, z_{1+h}, z_{1+2h}, ...\}$, $\{z_2, z_{2+h}, z_{2+2h}, ...\}$, $\{z_3, z_{3+h}, z_{3+2h}, ...\}$, and $\{z_4, z_{4+h}, z_{4+2h}, ...\}$.

* indicates rejection at 5% significance level with Bonferroni bounds.

Table 8. Berkowitz (2001) Likelihood Ratio test at $h = 1$

Output Growth				Inflation			
Variable	$\mu = 0, \sigma = 1$	$\rho = 0$	joint	Variable	$\mu = 0, \sigma = 1$	$\rho = 0$	joint
rbndm@us	0.00 *	0.14	0.00 *	rbndm@us	0.00 *	0.82	0.00 *
rbndl@us	0.00 *	0.15	0.00 *	rbndl@us	0.00 *	0.86	0.00 *
rspread@us	0.00 *	0.30	0.00 *	rspread@us	0.00 *	0.84	0.00 *
stockp@us	0.00 *	0.67	0.00 *	stockp@us	0.00 *	0.88	0.00 *
exrate@us	0.29	0.53	0.39	exrate@us	0.00 *	0.64	0.00 *
rrovnght@us	0.00 *	0.33	0.00 *	rrovnght@us	0.00 *	0.64	0.00 *
rrtbill@us	0.00 *	0.33	0.00 *	rrtbill@us	0.00 *	0.88	0.00 *
rrbnds@us	0.00 *	0.33	0.01 *	rrbnds@us	0.00 *	0.92	0.00 *
rrbndm@cn	0.00 *	0.15	0.00 *	rrbndm@cn	0.00 *	0.49	0.00 *
rrbndl@us	0.00 *	0.13	0.00 *	rrbndl@us	0.00 *	0.65	0.00 *
rstockp@us	0.00 *	0.62	0.00 *	rstockp@us	0.00 *	0.88	0.00 *
rexrate@us	0.29	0.53	0.39	rexrate@us	0.00 *	0.64	0.00 *
ip@us	0.03 *	0.30	0.04 *	rgdp@us	0.00 *	0.23	0.00 *
capu@us	0.00 *	0.60	0.01 *	ip@us	0.00 *	0.60	0.00 *
emp@us	0.02 *	0.48	0.04 *	capu@us	0.00 *	0.71	0.00 *
unemp@us	0.00 *	0.67	0.00 *	emp@us	0.00 *	0.57	0.00 *
pgdp@us	0.00 *	0.26	0.00 *	unemp@us	0.00 *	0.43	0.00 *
cpi@us	0.00 *	0.22	0.00 *	cpi@us	0.00 *	0.78	0.00 *
ppi@us	0.00 *	0.76	0.00 *	ppi@us	0.00 *	0.48	0.00 *
earn@us	0.00 *	0.95	0.00 *	earn@us	0.00 *	0.99	0.00 *
mon0@us	0.00 *	0.80	0.01 *	mon0@us	0.00 *	0.65	0.00 *
mon1@us	0.00 *	0.21	0.00 *	mon1@us	0.00 *	0.95	0.00 *
mon2@us	0.00 *	0.22	0.00 *	mon2@us	0.00 *	0.89	0.00 *
mon3@us	0.00 *	0.86	0.00 *	mon3@us	0.00 *	0.78	0.00 *
rmon0@us	0.00 *	0.69	0.00 *	rmon0@us	0.00 *	0.44	0.00 *
rmon1@us	0.00 *	0.08	0.00 *	rmon1@us	0.00 *	0.87	0.00 *
rmon2@us	0.00 *	0.34	0.00 *	rmon2@us	0.00 *	0.93	0.00 *
rmon3@us	0.01 *	0.99	0.03 *	rmon3@us	0.00 *	0.62	0.10
Simple Average	0.48	0.93	0.69	Simple Average	0.00 *	0.62	0.00 *
BMA-OLS	0.00 *	0.91	0.00 *	BMA-OLS	0.00 *	0.72	0.00 *
BMA	0.02 *	0.11	0.03 *	BMA	0.10	0.40	0.10
BMA-MC3	0.00 *	0.00 *	0.00 *	BMA-MC3	0.00 *	0.00 *	0.00 *
Factor	0.00 *	0.43	0.00 *	Factor	0.00 *	0.71	0.00 *
BVAR	0.00 *	0.05	0.00 *	BVAR	0.00 *	0.00 *	0.00 *

Notes: The table reports p-values of Berkowitz LR test under various null hypotheses. * indicates rejection at 5% significance level.

Table 9. Berkowitz (2001) Likelihood Ratio test at $h = 4$

Output Growth				Inflation			
Variable	$\mu = 0, \sigma = 1$	$\rho = 0$	joint	Variable	$\mu = 0, \sigma = 1$	$\rho = 0$	joint
rgdp@us	0.03	0.62	0.06	pgdp@us	0.30	0.28	0.35
rovnght@us	0.26	0.02	0.08	rovnght@us	0.00 *	0.13	0.01 *
rtbill@us	0.39	0.00 *	0.01 *	rtbill@us	0.00 *	0.20	0.00 *
rbnds@us	0.42	0.02	0.08	rbnds@us	0.00 *	0.21	0.00 *
rbndm@us	0.83	0.10	0.44	rbndm@us	0.00 *	0.22	0.02
rbndl@us	0.72	0.11	0.46	rbndl@us	0.01 *	0.17	0.02
rspread@us	0.09	0.30	0.20	rspread@us	0.16	0.22	0.21
stockp@us	0.24	0.36	0.44	stockp@us	0.08	0.26	0.16
exrate@us	0.41	0.55	0.57	exrate@us	0.14	0.11	0.31
rrovnght@us	0.25	0.24	0.39	rrovnght@us	0.39	0.33	0.44
rrtbill@us	0.80	0.21	0.65	rrtbill@us	0.24	0.33	0.30
rrbnds@us	0.50	0.18	0.55	rrbnds@us	0.17	0.16	0.15
rrbndm@cn	0.26	0.12	0.23	rrbndm@cn	0.10	0.22	0.11
rrbndl@us	0.27	0.10	0.09	rrbndl@us	0.07	0.22	0.08
rstockp@us	0.22	0.29	0.41	rstockp@us	0.10	0.20	0.18
rexrate@us	0.41	0.55	0.57	rexrate@us	0.14	0.11	0.31
ip@us	0.04	0.69	0.09	rgdp@us	0.15	0.23	0.26
capu@us	0.39	0.38	0.53	ip@us	0.28	0.40	0.38
emp@us	0.03	0.65	0.06	capu@us	0.19	0.04	0.08
unemp@us	0.11	0.17	0.10	emp@us	0.17	0.41	0.30
pgdp@us	0.46	0.36	0.55	unemp@us	0.12	0.17	0.22
cpi@us	0.08	0.02	0.04	cpi@us	0.03	0.13	0.05
ppi@us	0.17	0.45	0.21	ppi@us	0.11	0.08	0.13
earn@us	0.43	0.53	0.54	earn@us	0.04	0.17	0.10
mon0@us	0.44	0.36	0.51	mon0@us	0.13	0.18	0.10
mon1@us	0.15	0.48	0.24	mon1@us	0.02	0.20	0.05
mon2@us	0.35	0.28	0.39	mon2@us	0.03	0.18	0.05
mon3@us	0.11	0.11	0.23	mon3@us	0.02	0.33	0.06
rmon0@us	0.07	0.23	0.15	rmon0@us	0.25	0.12	0.11
rmon1@us	0.18	0.02	0.06	rmon1@us	0.09	0.11	0.08
rmon2@us	0.37	0.25	0.39	rmon2@us	0.07	0.10	0.02
rmon3@us	0.46	0.43	0.57	rmon3@us	0.50	0.15	0.55
Simple Average	0.02	0.50	0.04	Simple Average	0.25	0.18	0.31
BMA-OLS	0.06	0.30	0.06	BMA-OLS	0.00 *	0.07	0.02
BMA	0.32	0.24	0.40	BMA	0.34	0.11	0.22
BMA-MC3	0.00 *	0.00 *	0.00 *	BMA-MC3	0.00 *	0.00 *	0.00 *
Factor	0.23	0.58	0.34	Factor	0.02	0.11	0.05
BVAR	0.00 *	0.02	0.00 *	BVAR	0.01 *	0.12	0.01 *

Notes: The table reports minimum p-values of the Berkowitz LR test under different null hypotheses based on four subsets $\{z_1, z_{1+h}, z_{1+2h}, ...\}$, $\{z_2, z_{2+h}, z_{2+2h}, ...\}$, $\{z_3, z_{3+h}, z_{3+2h}, ...\}$, and $\{z_4, z_{4+h}, z_{4+2h}, ...\}$.

* indicates rejection at 5% significance level with Bonferroni bounds.

Table 10. Doornik and Hansen (2008) test

Output Growth			Inflation		
Variable	$h = 1$	$h = 4$	Variable	$h = 1$	$h = 4$
rgdp@us	0.18	0.05	pgdp@us	0.19	0.01 *
rovnght@us	0.56	0.05	rovnght@us	0.74	0.01 *
rtbill@us	0.42	0.07	rtbill@us	0.51	0.09
rbnds@us	0.48	0.05	rbnds@us	0.64	0.04
rbndm@us	0.31	0.02	rbndm@us	0.55	0.01 *
rbndl@us	0.54	0.04	rbndl@us	0.46	0.01 *
rspread@us	0.32	0.03	rspread@us	0.48	0.05
stockp@us	0.46	0.02	stockp@us	0.23	0.02
exrate@us	0.20	0.01 *	exrate@us	0.34	0.05
rrovnght@us	0.29	0.00 *	rrovnght@us	0.35	0.04
rrtbill@us	0.25	0.02	rrtbill@us	0.28	0.01 *
rrbnds@us	0.37	0.01 *	rrbnds@us	0.29	0.01 *
rrbndm@cn	0.30	0.07	rrbndm@cn	0.25	0.02
rrbndl@us	0.38	0.04	rrbndl@us	0.39	0.03
rstockp@us	0.39	0.02	rstockp@us	0.22	0.05
rexrate@us	0.20	0.01 *	rexrate@us	0.34	0.05
ip@us	0.05	0.04	rgdp@us	0.35	0.04
capu@us	0.13	0.03	ip@us	0.33	0.04
emp@us	0.36	0.09	capu@us	0.35	0.02
unemp@us	0.37	0.02	emp@us	0.54	0.29
pgdp@us	0.33	0.21	unemp@us	0.56	0.03
cpi@us	0.37	0.06	cpi@us	0.52	0.08
ppi@us	0.25	0.22	ppi@us	0.21	0.02
earn@us	0.67	0.21	earn@us	0.50	0.04
mon0@us	0.15	0.04	mon0@us	0.17	0.07
mon1@us	0.42	0.08	mon1@us	0.16	0.09
mon2@us	0.44	0.08	mon2@us	0.19	0.01 *
mon3@us	0.16	0.02	mon3@us	0.44	0.04
rmon0@us	0.22	0.09	rmon0@us	0.43	0.01 *
rmon1@us	0.51	0.17	rmon1@us	0.16	0.00 *
rmon2@us	0.43	0.06	rmon2@us	0.32	0.03
rmon3@us	0.30	0.02	rmon3@us	0.23	0.11
Simple Average	0.03 *	0.02	Simple Average	0.19	0.00 *
BMA-OLS	0.39	0.20	BMA-OLS	0.83	0.04
BMA	0.05	0.13	BMA	0.20	0.01 *
BMA-MC3	0.00 *	0.15	BMA-MC3	0.04 *	0.29
Factor	0.28	0.12	Factor	0.59	0.12
BVAR	0.00 *	0.00 *	BVAR	0.00 *	0.00 *

Notes: The table reports p-values for $h = 1$ and minimum p-values for $h = 4$ of the Doornik-Hansen test.

* indicates rejection at 5% significance level (with Bonferroni bounds for $h = 4$ case).

Table 11, Panel A. Summary of the Results (h=1)

	Uniformity		Un-correlation		Stability		Unif. & Uncorr. (Berkowitz)			DH
	KS	AD	z_{t+h}	z^2_{t+h}	z_{t+h}	z^2_{t+h}	$\mu=0,\sigma=1$	$\rho=0$	joint	
	Output Growth									
AR	yes	yes	yes	no	yes	yes	no	yes	no	yes
ADL	30/31	13/31	30/31	20/31	30/31	29/31	2/31	31/31	2/31	31/31
Simple Av.	yes	yes	yes	yes	yes	yes	yes	yes	yes	no
BMA-OLS	yes	no	yes	yes	yes	yes	no	yes	no	yes
BMA	yes	no	yes	yes	yes	yes	no	yes	no	yes
BMA-MC3	no	no	no	no	no	no	no	no	no	no
Factor	yes	yes	yes	yes	yes	yes	no	yes	no	yes
BVAR	no	no	yes	no	yes	yes	no	yes	no	no
	Inflation									
AR	yes	no	yes	no	yes	yes	no	yes	no	yes
ADL	31/31	0/31	31/31	4/31	31/31	31/31	0/31	31/31	1/31	31/31
Simple Av.	yes	yes	yes	no	yes	yes	no	yes	no	yes
BMA-OLS	yes	no	yes	yes	yes	yes	no	yes	no	yes
BMA	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
BMA-MC3	no	no	no	no	no	no	no	no	no	no
Factor	yes	no	yes	no	yes	yes	no	yes	no	yes
BVAR	no	no	no	no	yes	yes	no	no	no	no

Table 11, Panel B. Summary of the Results (h=4)

	Uniformity		Un-correlation		Stability		Unif. & Uncorr. (Berkowitz)			DH
	KS	AD	z_{t+h}	z^2_{t+h}	z_{t+h}	z^2_{t+h}	$\mu=0,\sigma=1$	$\rho=0$	joint	
	Output Growth									
AR	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
ADL	31/31	29/31	29/31	30/31	26/31	15/31	31/31	30/31	30/31	27/31
Simple Av.	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
BMA-OLS	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
BMA	yes	yes	yes	yes	yes	no	yes	yes	yes	yes
BMA-MC3	no	no	no	yes	no	no	no	no	no	yes
Factor	yes	yes	yes	yes	yes	yes	yes	yes	yes	yes
BVAR	no	no	yes	yes	no	no	no	yes	no	no
	Inflation									
AR	yes	yes	yes	no	yes	yes	yes	yes	yes	no
ADL	30/31	24/31	31/31	20/31	26/31	26/31	26/31	31/31	28/31	23/31
Simple Av.	yes	yes	yes	no	yes	yes	yes	yes	yes	no
BMA-OLS	no	no	yes	yes	yes	no	no	yes	yes	yes
BMA	yes	yes	yes	yes	yes	yes	yes	yes	yes	no
BMA-MC3	yes	no	no	no	no	no	no	no	no	yes
Factor	yes	yes	yes	no	yes	yes	yes	yes	yes	yes
BVAR	yes	no	yes	no	yes	yes	no	yes	no	no

Notes: The table shows whether the specific test indicated by the column provides statistical evidence in support of the proper calibration of PITs implied by the models in each row, (e.g. "yes" in the uniformity column means the test does not reject uniformity). For the ADL models, we report how many specifications (across the various predictors) are not rejected by the specified test. Rejections are at 5% significance level as listed in Tables 2-10.

Figure 1, Panel A: PITs for ADL Models of Output Growth at $h = 1$

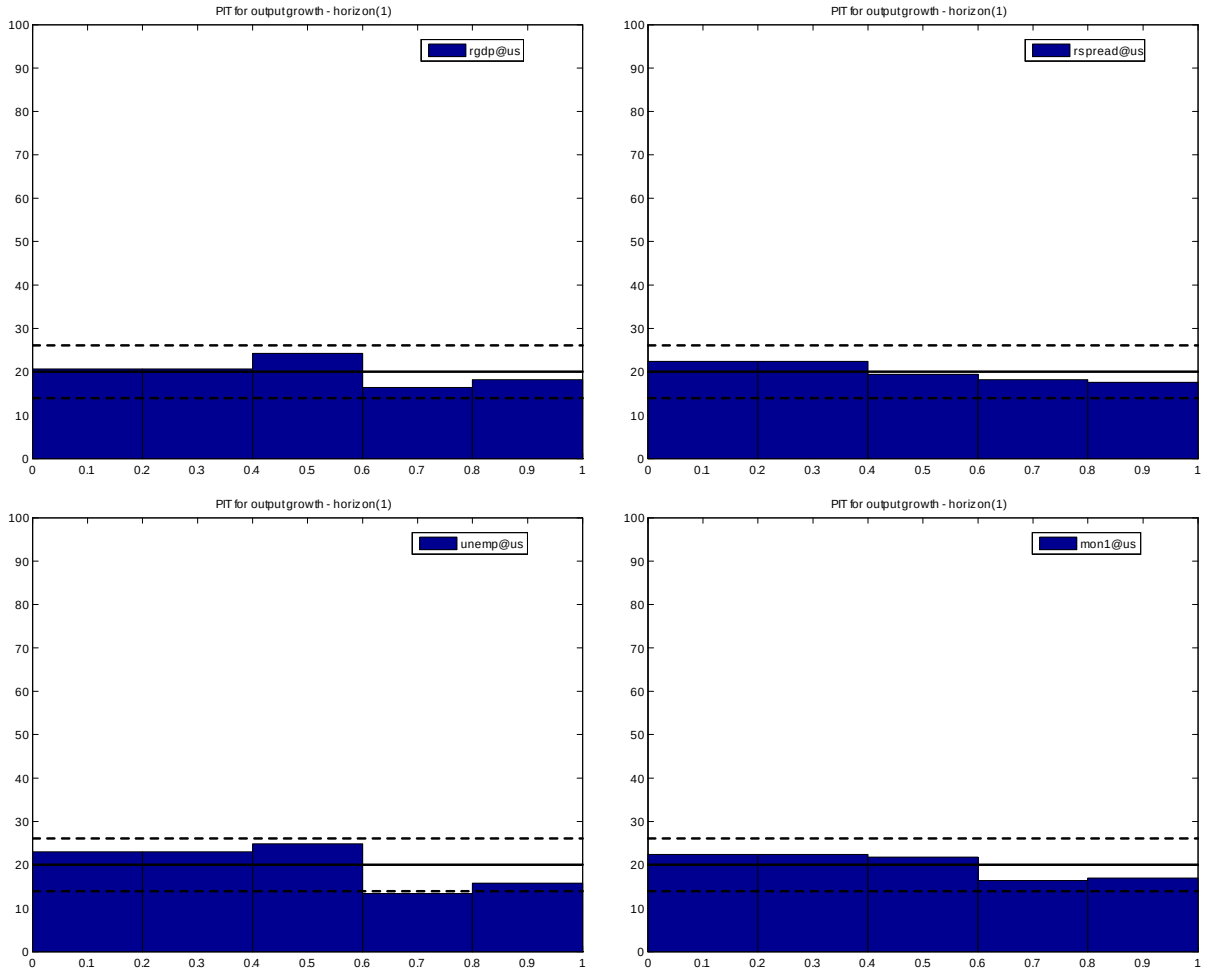
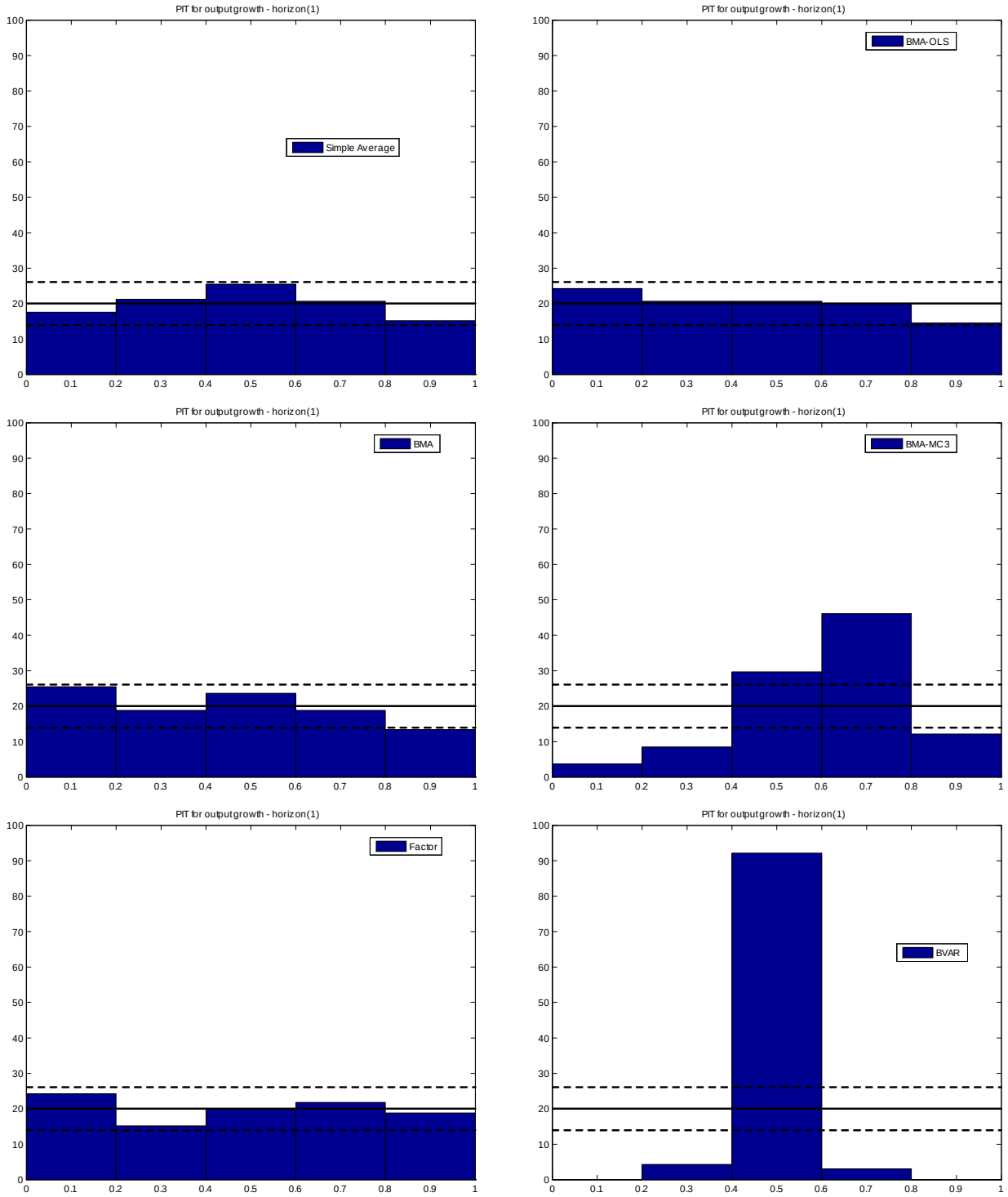


Figure 1, Panel B: PITs for Models Combining Large Data Sets of Output Growth at $h = 1$



Notes: The histograms depict the empirical distributions of the PITs. Solid line represents the number of draws that are expected to be in each bin under U(0,1) distribution. The dashed lines represent the 95% confidence interval constructed under the normal approximation of a binomial distribution.

Figure 2, Panel A: PITs for ADL Models of Inflation at $h = 1$

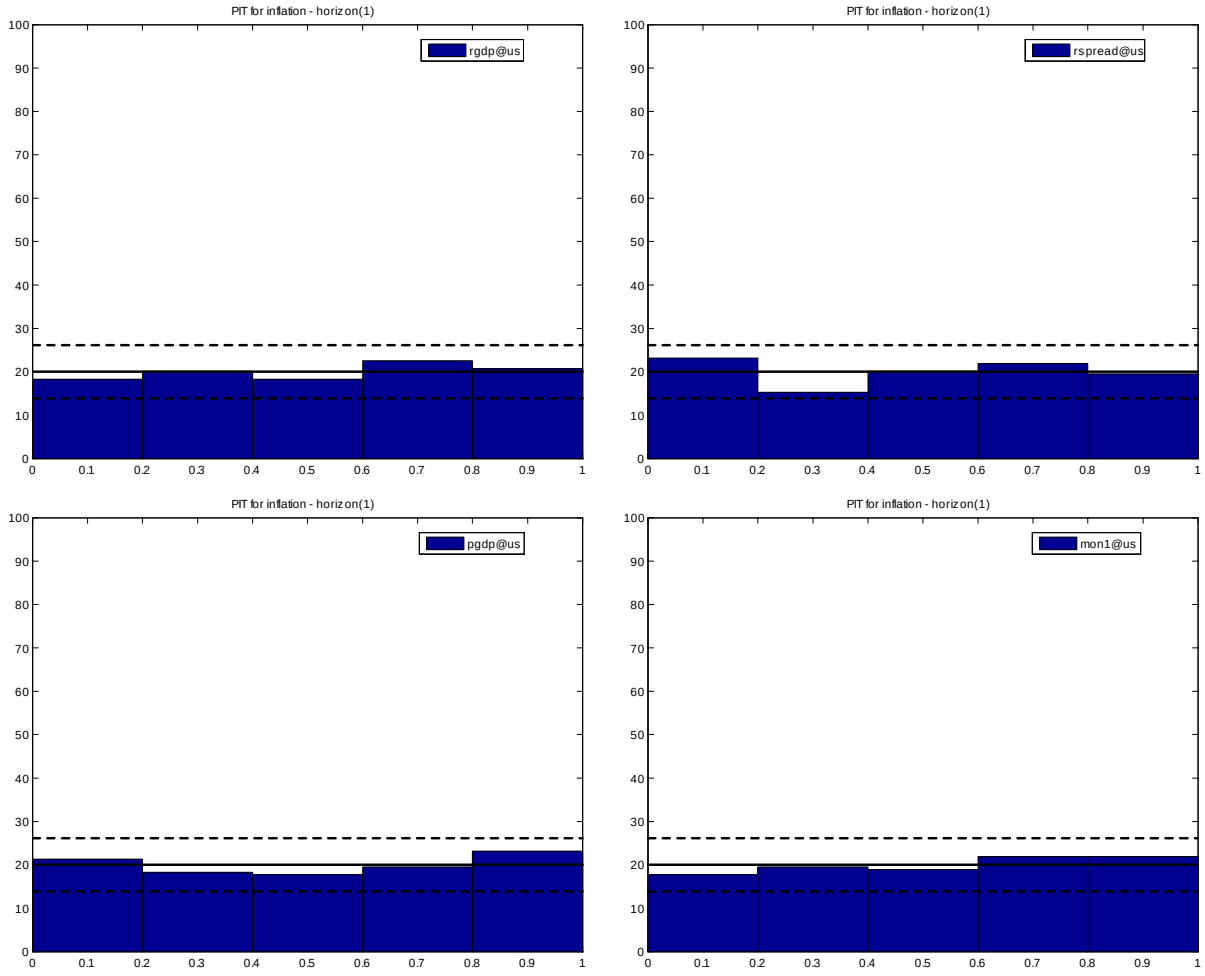
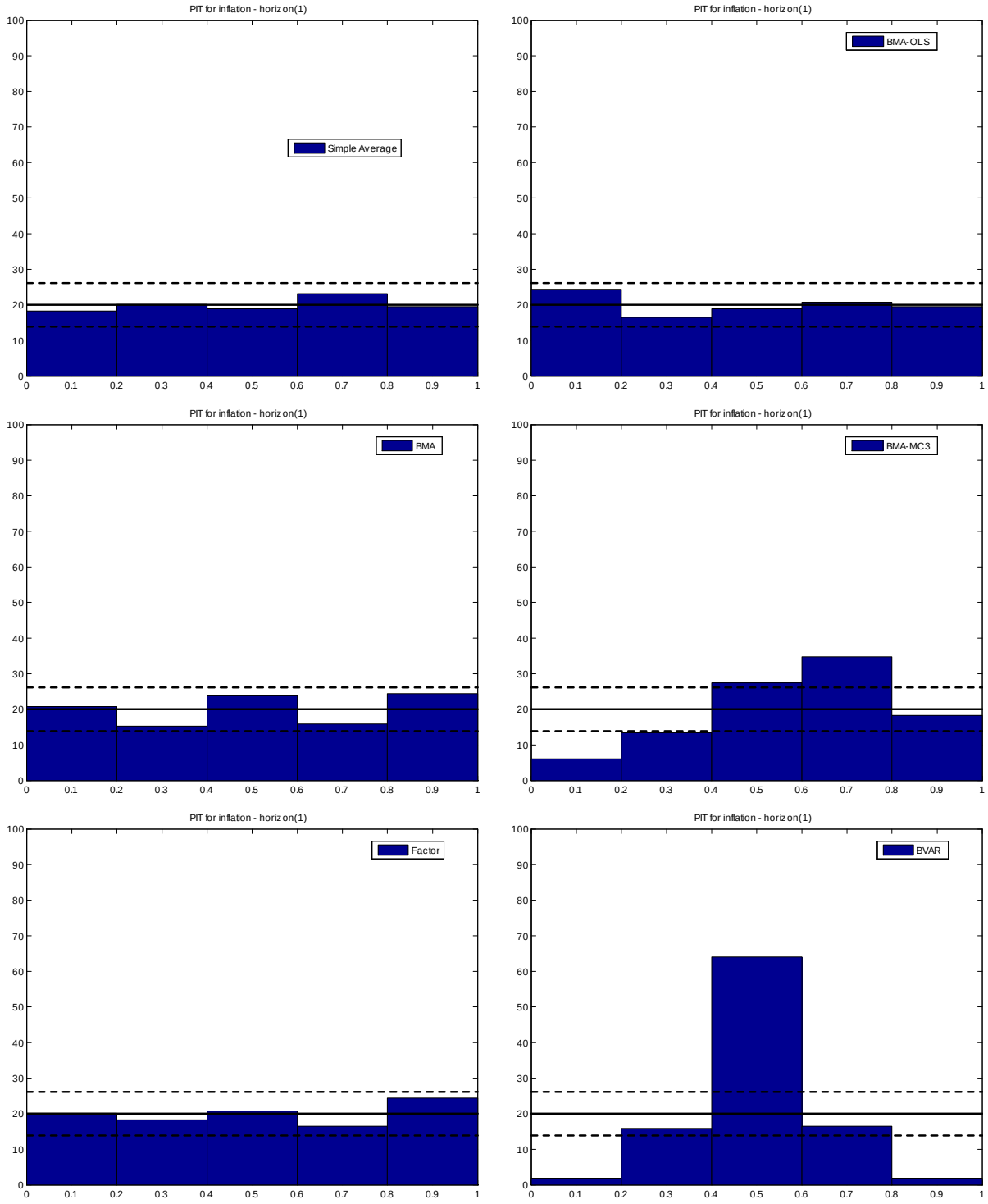


Figure 2, Panel B: PITs for Models Combining Large Data Sets of Inflation at $h = 1$



Notes: The histograms depict the empirical distributions of the PITs. Solid line represents the number of draws that are expected to be in each bin under U(0,1) distribution. The dashed lines represent the 95% confidence interval constructed under the normal approximation of a binomial distribution.