International Journal of Forecasting 30 (2014) 662–682

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set

Barbara Rossi [a,*], Tatevik Sekhposyan [b]

[a] ICREA-UPF, Barcelona GSE, and CREI, Carrer Ramon Trias Fargas 25-27, Mercè Rodoreda bldg., 08005 Barcelona, Spain
[b] Bank of Canada, 234 Wellington Street, Ottawa, ON, K1A 0G9, Canada

ARTICLE INFO

Keywords:
Predictive density evaluation
Structural change
Output growth forecasts
Inflation forecasts

ABSTRACT

We evaluate conditional predictive densities for US output growth and inflation using a number of commonly-used forecasting models that rely on large numbers 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 cause point forecasts to either improve or deteriorate, 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 approximated correctly by a normal density: the simple, equal average when predicting output growth, and the Bayesian model average when predicting inflation.

1. Introduction

Forecasts are traditionally used to evaluate models' performances. In most cases, forecasts are judged as good or otherwise based mainly on the models' (median or mean) point forecasts. For example, Stock and Watson (2003) conducted an extensive evaluation of a large data set of predictors of US output growth and inflation, focusing on point forecasts; while Banerjee and Marcellino (2006), Banerjee, Marcellino, and Masten (2005) and Marcellino, Stock, and Watson (2003) conducted similarly broad analyses for the Euro area. Furthermore, Rossi and Sekhposyan (2010) 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 determine the correct specification of the uncertainty around models' point forecasts. For example, central banks are increasingly concerned about the 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 used extensively in the literature for forecasting output growth and inflation (and seemingly doing a good job, according to their point forecasts), and investigate whether their predictive densities are calibrated correctly by the commonly-used normal approximation (see Stock & Watson, 2002). We use the Probability Integral Transform (PIT) technique, which was originally introduced by Rosenblatt (1952), and more recently has been proposed by Diebold, Gunther, and Tay (1998) for evaluating the correct specifications of predictive densities. Corradi and Swanson (2006) provide a comprehensive recent overview of tests for predictive density evaluation, and Garratt, Lee, Pesaran, and Shin (2003) and Granger and Pesaran (2000) complement the discussion further. The differences between this paper and those in the previous literature are the fact that we operate in a data-rich environment using the extensive data set of Stock and Watson (2003), and 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

* Corresponding author.
E-mail addresses: barbara.rossi@upf.edu (B. Rossi), tsekhposyan@bankofcanada.ca (T. Sekhposyan).

0169-2070/$ – see front matter
http://dx.doi.org/10.1016/j.ijforecast.2013.03.005

not only predictive densities based on autoregressive distributed lag (ADL) models with several predictors considered one at a time (as did Stock & Watson, 2003), but also forecast combinations. We include predictive density combinations with either equal weights or weights which are equal to the posterior probabilities of the models. In addition, we also consider several different estimation techniques: we combine models estimated by OLS with those from Bayesian shrinkage methods and a posterior simulator algorithm that samples models from the model space with the 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 determine 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 by Diebold et al. (1998) and Diebold, Tay, and Wallis (1999), as well as the Kolmogorov–Smirnov and Anderson–Darling tests. We also consider tests based on the inverse normal transformation of the PIT, including 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 the 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 performances of ADL models across the various tests depend 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, particularly 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 at either the estimation or 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 has 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 & Watson, 2003; Timmermann, 2006; Wright, 2009). When testing the appropriateness of the normal distribution, we find that this also extends to density forecasts.

In more detail, based on both the Kolmogorov–Smirnov and Anderson–Darling tests, we find more pervasive evidence against uniformity for the 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 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, such 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. This is a result that holds in general based on variety of tests except the Doornik and Hansen's (2008) test. 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, the predictive densities of simple averaging and BMA models are among the best calibrated, in spite of the target variable which we consider. The occasional failings are associated mainly 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 with a lack of uniformity of the PITs at the one-quarter-ahead forecast horizon. Similarly, the BMA also performs fairly well for output growth, although it fails uniformity for the one-quarter-ahead forecast horizon, and stability for one-year-ahead.

An analysis which is similar in spirit to the one considered in this paper is that of Clements and Smith (2000). However, there are several differences between our work and theirs. First, they focus only on forecasting output growth and unemployment, and do not consider forecasts of inflation, which is another important variable for which we are interested in the predictive density. 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 (like most papers that evaluate density forecasts, starting from Diebold et al., 1998) focuses on testing the uniformity and uncorrelatedness of the PITs, whereas we also formally test the hypothesis of identical distributions over time.

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

[1] Note that our focus throughout this paper is on testing for 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 the correct specification is not random over time, and may signal misspecification in the dynamics of the underlying models.

Our paper also differs from those of Jore, Mitchell, and Vahey (2010) and Manzan and Zerom (2013). Jore et al. (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 the 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 closest 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 was undertaken only recently by Manzan and Zerom (2013), who compared the 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, and Section 3 discusses the set of forecasting models. Section 4 then describes the data and the empirical results, and 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 $t = R, \ldots, T$ over a window of R observations, including data indexed from $t - R + 1$ 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 of resulting in properly calibrated densities—see Clark (2011) and Jore et al. (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 et al. (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 et al. (1998), if the proposed predictive density is consistent with the true predictive density, then, for $h = 1$, the density of $\{z_{t+h}\}_{t=R}^{T}$ will be an independent and identically distributed (i.i.d.) Uniform (0, 1), and its cumulative distribution function will be the 45° line. When $h > 1$, independence is violated by construction, even if the models are correctly specified, since serial correlation of order $(h - 1)$ is built into the multi-step-ahead density forecasts by construction. One recommendation given by Clements and Smith (2000) and Diebold et al. (1998), among others, was 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 to detect the absence of serial correlation in the PITs; the rest of the tests aim to detect violations of uniformity (at times jointly 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 also implement Andrews' (1993) QLR test for evaluating 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. On the other hand, Berkowitz (2001) 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 to the PITs directly, at least in finite samples. In what follows, we provide a detailed discussion of the characteristics of each of the tests we implement.

2.1. Tests on the PIT

2.1.1. Diebold et al. (1998) test

Diebold et al. (1998) rely mainly on a graphical assessment of the uniformity and independence properties that

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

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

characterize the PITs of correctly specified predictive distributions. Following Diebold et al. (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 functions of various powers of the PITs. Specifically, we follow Diebold et al. (1998) in deriving confidence intervals for the number of observations which fall into any one bin, in order to assess the uniformity of the PITs statistically; 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 show the fractions of the PITs which fall into each bin. If the PITs are indeed i.i.d. uniform, then each bin will 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 benefit 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.

2.1.2. Tests of uniformity

We test whether the PIT is uniform using the Kolmogorov–Smirnov and Anderson–Darling tests. The former measures the difference between the empirical distribution of the PITs, $\widehat{\Phi}_{t+h}(y_{t+h}|\Im_t)$, and the cumulative distribution of a uniform distribution, $r \in (0, 1)$ (i.e., the 45° line). The Anderson–Darling test is a special type of the Cramér–von Mises test, which places more weight on the deviations in the tails of the empirical distribution, rather than weighting all of 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)

$$\mathrm{KS} = \sqrt{P} \max_{j=1,\ldots,P} \max\{|z_j^\dagger - j/P|, |z_j^\dagger - (j-1)/P|\}. \quad (2)$$

ii. Anderson–Darling (AD, Anderson & Darling, 1952, 1954)

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

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

2.1.3. 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

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

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 (see also Hayashi, 2000, p. 144).

2.1.4. Tests of identical distribution

Complementing the empirical evidence, we also consider tests for identical distributions. If z_{t+h} were distributed identically 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 typically been used in the forecasting literature for judging whether predictors' Granger-causality is stable over time: see Stock and Watson (2003). Here, we are concerned with the question of 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 the following regressions:[6]

$$z_{t+h} = \alpha_{1,t} + \varepsilon_{1,t+h} \quad (5)$$

$$z_{t+h}^2 = \alpha_{2,t} + \varepsilon_{2,t+h}. \quad (6)$$

2.2. Tests on the inverse normal of the PIT

Berkowitz (2001, Proposition 1) shows that if the PIT is i.i.d. $U(0, 1)$, then the inverse standard normal transformation of the PIT is an i.i.d. 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.

2.2.1. Berkowitz's (2001) test

Berkowitz (2001) proposes a joint test for a zero mean, unit variance, and independence in $\widetilde{z}_{t+h}$, against an autoregressive alternative with a mean and a variance which may differ from 0 and 1, respectively. That is, we jointly test whether $\mu = 0, \sigma = 1$ and $\rho = 0$ in the regression:

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

where $\varepsilon_{t+h} \sim (0, \sigma^2)$.[7] This test is implemented as a likelihood ratio (LR) test, which has an asymptotic $\chi^2(3)$ distribution under the null hypothesis described above. One could also test a subset of the hypotheses in this setting;

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

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

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

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

Table 1
Description of the 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	Δ ln	59:Q1–10:Q3	SP500	US Share Prices: S&P 500	F
exrate@us	Δ ln	73:M1–10:M9	111…NELZF…	NEER from UCL	I
Real activity					
rgdp@us	Δ ln	59:Q1–10:Q3	GDPC96	Real GDP, sa	F
ip@us	Δ ln	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	Δ ln	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	Δ ln	59:Q1–10:Q3	GDPDEF	GDP Deflator, sa	F
cpi@us	Δ ln	59:M1–10:M9	CPIAUCSL	CPI: Urban, All items, sa	F
ppi@us	Δ ln	59:M1–10:M9	PPIACO	Producer Price Index, nsa	F
earn@us	Δ ln	59:M1–10:M9	AHEMAN	Hourly Earnings: Man., nsa	F
Money					
mon0@us	Δ ln	59:M1–10:M9	AMBSL	Monetary Base, sa	F
mon1@us	Δ ln	59:M1–10:M9	M1SL	Money: M1, sa	F
mon2@us	Δ ln	59:M1–10:M9	M2SL	Money: M2, sa	F
mon3@us	Δ ln	59:M1–06:M2	M3SL	Money: M3, sa	F

Notes: The 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 by a prefix "r", it indicates a real variable adjusted by either the CPI (stock variables) or CPI inflation (flow variables). The interest rate spread is calculated as the difference between "rbndl" and "rovnght".

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 those under the PIT framework is that the test of 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 to the PITs directly, at least in small samples; the limitation is that they only detect violations of normality through the first two moments, not higher moments, whereas PIT-based tests can detect any departure from uniformity.

2.2.2. *Doornik and Hansen's (2008) test*

Doornik and Hansen (2008) proposed testing the normality of $\widetilde{z}_{t+h}$ using a test of skewness and kurtosis which has good small sample properties. The test is based on the sum of the squares of transformed measures of skewness and kurtosis, and has a $\chi^2(2)$ asymptotic distribution under the null of i.i.d. normality (i.e., an absence of skewness and kurtosis).

3. Forecasting models

All of the models which we consider are estimated using the Stock and Watson (2003) database, collected at the quarterly frequency and updated its last available value in January 2011. The variables are asset prices, measures of real economic activity, wages and prices, and money. We follow Stock and Watson (2003) and transform the data in order to eliminate stochastic or deterministic trends, as well as seasonality. For example, all of 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 later starting dates 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.

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 the 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 for $t = 1, \ldots, T$ is:

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

where the dependent variable is either $Y^h_{t+h} = (400/h)\ln(\text{RGDP}_{t+h}/\text{RGDP}_t)$ or $Y^h_{t+h}=(400/h)\ln(\text{PGDP}_{t+h}/\text{PGDP}_t) - 400\ln(\text{PGDP}_t/\text{PGDP}_{t-1})$, where RGDP_{t+h} and PGDP_{t+h} are the real GDP and the 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 kth variable, for $k = 1,\ldots,K$. 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 $Y_t = 400\ln(\text{RGDP}_t/\text{RGDP}_{t-1})$, or the period t change in inflation, that is $Y_t = 400\ln(\text{PGDP}_t/\text{PGDP}_{t-1}) - 400\ln(\text{PGDP}_{t-1}/\text{PGDP}_{t-2})$.[9] Furthermore, the error term u_{t+h} is assumed to be distributed normally, $N(0,\sigma^2)$. We consider $h = 1, 4$ corresponding to the 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 numbers of lags (p and q) recursively using the BIC, first selecting the lag length for the autoregressive component, then augmenting it with an optimal lag length for the additional predictor. The PITs at a given time period $t+h$ are $\Phi_{t+h}(Y^h_{t+h}|(\hat\beta_0+\hat\beta_1(L)X_{t,k}+\hat\beta_2(L)Y_t),\hat\sigma^2)$, where $\hat{\cdot}$ indicates OLS estimates of the model's parameters, while Φ_{t+h} is the conditional CDF of the proposed normal distribution. When estimating $\hat\sigma^2$, we use HAC-robust variance estimates (Newey & West, 1987).[10]

As a particular case, we consider the autoregressive model, where we use only the lagged dependent variable for forecasting output growth and inflation. The PIT for the autoregressive model is $\Phi_{t+h}(Y^h_{t+h}|(\hat\beta_0+\hat\beta_2(L)Y_t),\hat\sigma^2)$, 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 Stock and Watson (2003, 2004) showed to perform well for point forecasts.[11] We follow Mitchell and Wallis (2011), and consider the predictive distribution of the combined model. Specifically, we estimate the ADL models in Eq. (8) for each of the regressors (one at a time), i.e. for $k = 1,\ldots,K$, 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 et al., 2010, Eq. (1)):

$$\Phi^c_{t+h} = \frac{1}{K}\sum_{k=1}^{K}\Phi_{t+h}(Y^h_{t+h}|\hat\beta_{k,0}+\hat\beta_{k,1}(L)X_{t,k} + \hat\beta_{k,2}(L)Y_t,\hat\sigma_k{}^2), \tag{9}$$

where the k subscripts in the conditional moments indicate that the parameters correspond to the kth 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 weights on all of 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. However, it differs from the simple model average in that 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 $D_t = \{Y_t, X_t, Y_{t-1}, X_{t-1},\ldots,Y_{t-R}, X_{t-R}\}$. In this case, the PIT is:

$$\Phi^{\text{BMA-OLS}}_{t+h} = \sum_{k=1}^{K}P_t(M_k|D_t)\Phi_{t+h}(Y^h_{t+h}|(\hat\beta_{k,0}+\hat\beta_{k,1}(L)X_{t,k} + \hat\beta_{k,2}(L)Y_t),(\hat\sigma_k)^2). \tag{10}$$

– *BMA.* The second version of BMA which we consider is the full Bayesian version, where the estimated parameters are posterior estimates rather than the OLS counterparts (which would be equivalent in the Bayesian framework to obtaining coefficients under a flat prior), and thus, are influenced by the choice of the prior distribution. Let $\tilde{\cdot}$ indicate estimates which are 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:

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

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

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

$$\tilde\beta_k|\tilde h_k \backsim N(\bar\beta_k,\tilde h_k^{-1}[g\tilde X_k'\tilde X_k]^{-1}), \tag{12}$$

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).

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

[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.

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

This is appropriate because the precision parameter is common to all models. Like Wright (2009), we assume $g = 1$, which places equal weights on the prior and the data in the posterior density of regression coefficients. In order to parameterize the prior further, we need values for $\bar{\beta}_k = [\bar{\beta}_{k,0}\ \bar{\beta}_{k,10}\ \cdots\ \bar{\beta}_{k,1p}\ \bar{\beta}_{k,20}\ \cdots\ \bar{\beta}_{k,2q}]'$ 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 the period 1947:Q1–1958:Q4, while the remaining coefficients are centered around zero.[13]

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

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

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

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

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

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

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

Furthermore, in this context, both the predictive density and the posterior model density have analytic solutions. The predictive density is given by

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

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 of the models are equally likely a priori, 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

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

Note that when $g = 0$, both the BMA-OLS and 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 lengths) are selected recursively via the BIC. We keep p and q fixed at their recursively selected levels for both the BMA-OLS and BMA specifications. Furthermore, as was noted by Wright (2009), the analytic 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 the cost of not being able to derive analytic solutions for the predictive densities and the 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.

(iii) *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 a multiplicity of models to be considered at a lower computational cost: in fact, it allows all regressors to enter the right hand side of the regression model (not just the autoregressive lags and the lags of only one additional economic variable). That is, MC3 is an algorithm that could help the researcher to 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 a lag length of q and one additional explanatory variable.[14]
- At step s, $s = \{1, 2, \ldots, S\}$, consider a new candidate model M^*, which is drawn randomly with equal probability from a set of models that includes: (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:

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

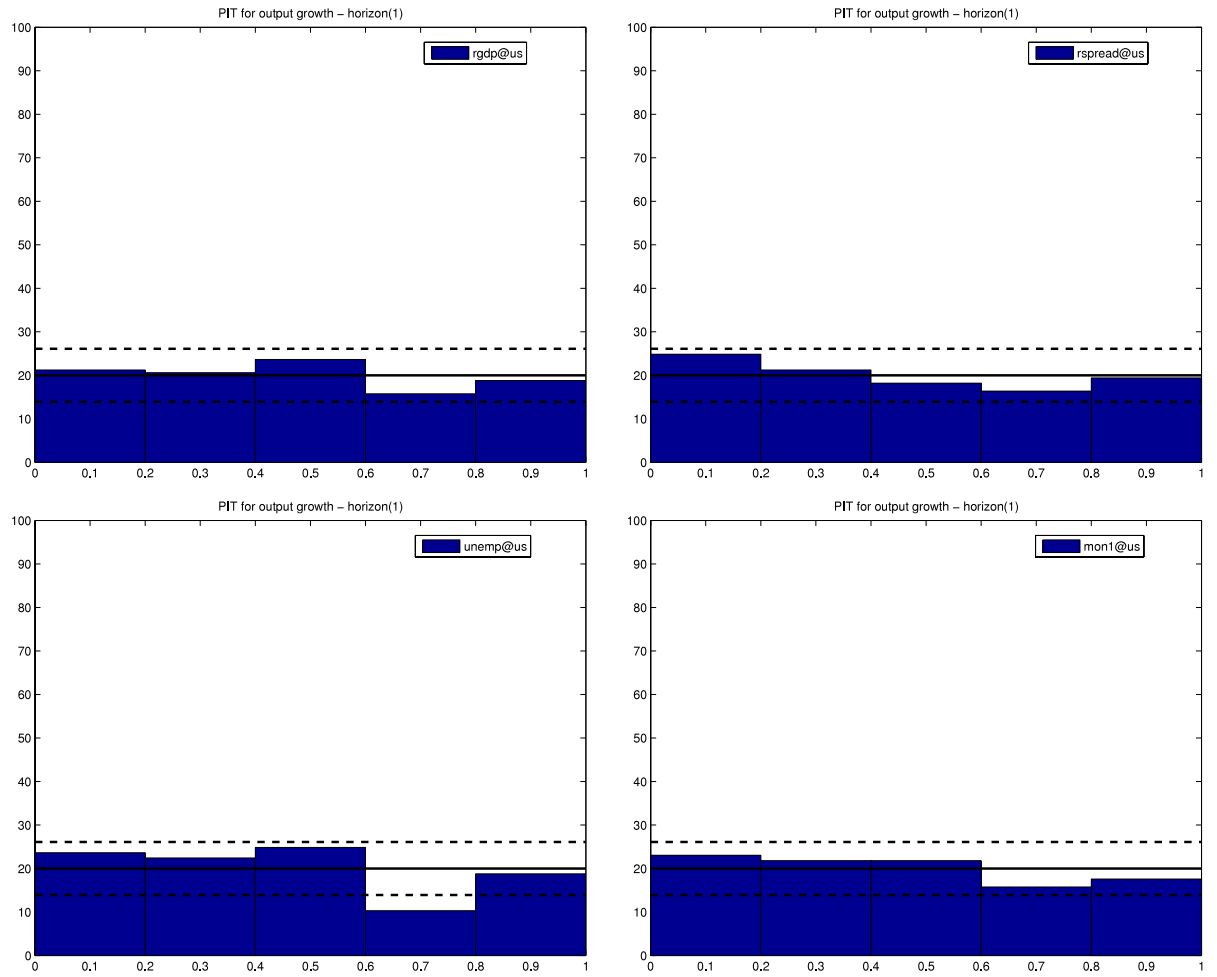
- We save $P_t(M_k|D_t)$ and $\tilde{\Phi}_{t+h}(Y_{t+h}|D_t, M_k)$ for accepted models.

Let $S = 10{,}000$ be the total number of draws, while $\bar{S} = 1000$ denotes the number of burn-in draws.[15]

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

[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] Burn-in draws are discarded in order to minimize the effect of the starting point on the simulation.



(a) Panel A: PITs for ADL models of output growth at $h = 1$.

Fig. 1. Notes: The histograms depict the empirical distributions of the PITs. Solid lines represent the numbers of draws that are expected to be in each bin under a $U(0, 1)$ distribution. Dashed lines represent the 95% confidence intervals constructed under the normal approximation of a binomial distribution.

The pooled predictive density is:

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

3.3. Models with principal components

Next, we consider a variant of the ADL model, Eq. (8), where, instead of considering all of the individual regressors one by one, we consider one model augmented with factors extracted from the set of all regressors. Specifically, we estimate a static factor model:[16]

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

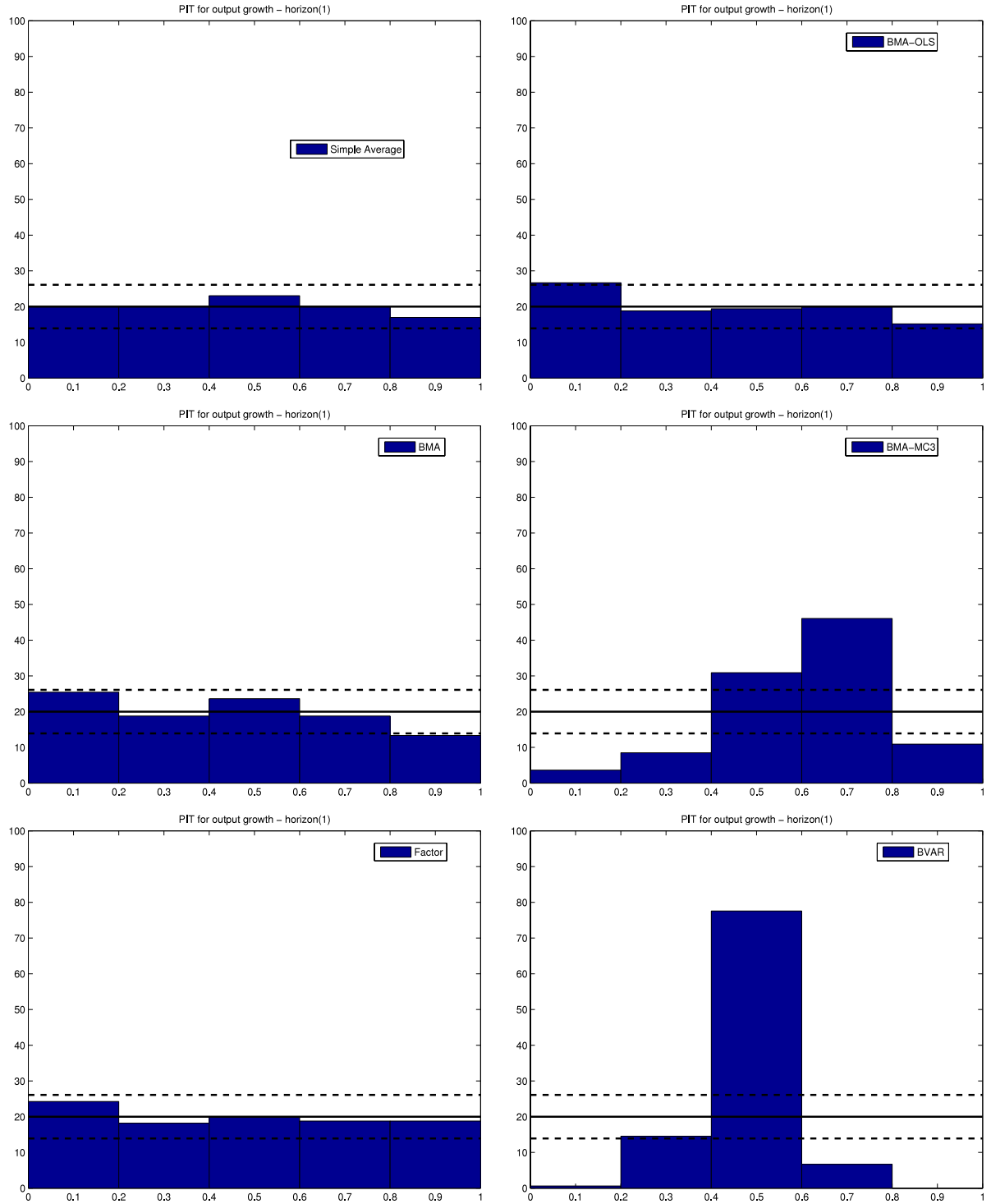
where $\hat{F}_t$ is the $(m \times 1)$ vector of estimated first m principal components of the K variables we consider in this paper. We select the number of factors m recursively over each rolling window R, such that the selected number of factors explains at least 60% of the variation contained in 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 $\Phi_{t+h}(Y_{t+h}^h|(\hat{\beta}_0 + \gamma\hat{F}_t + \hat{\beta}_2(L)Y_t), \hat{\sigma}^2)$, where $\hat{}$ indicates OLS estimates of the model parameters, while Φ_{t+h} is the conditional CDF of the proposed normal distribution, and $\hat{\sigma}^2$ is estimated by HAC.

3.4. Bayesian vector autoregressions

Finally, we consider a large scale Bayesian vector autoregression (BVAR) for modelling the joint dynamics

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

[17] Note that the dataset 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 it chooses a very large number of factors for our data set: for example, when the maximum number of factors allowed is 10, it chooses 10.



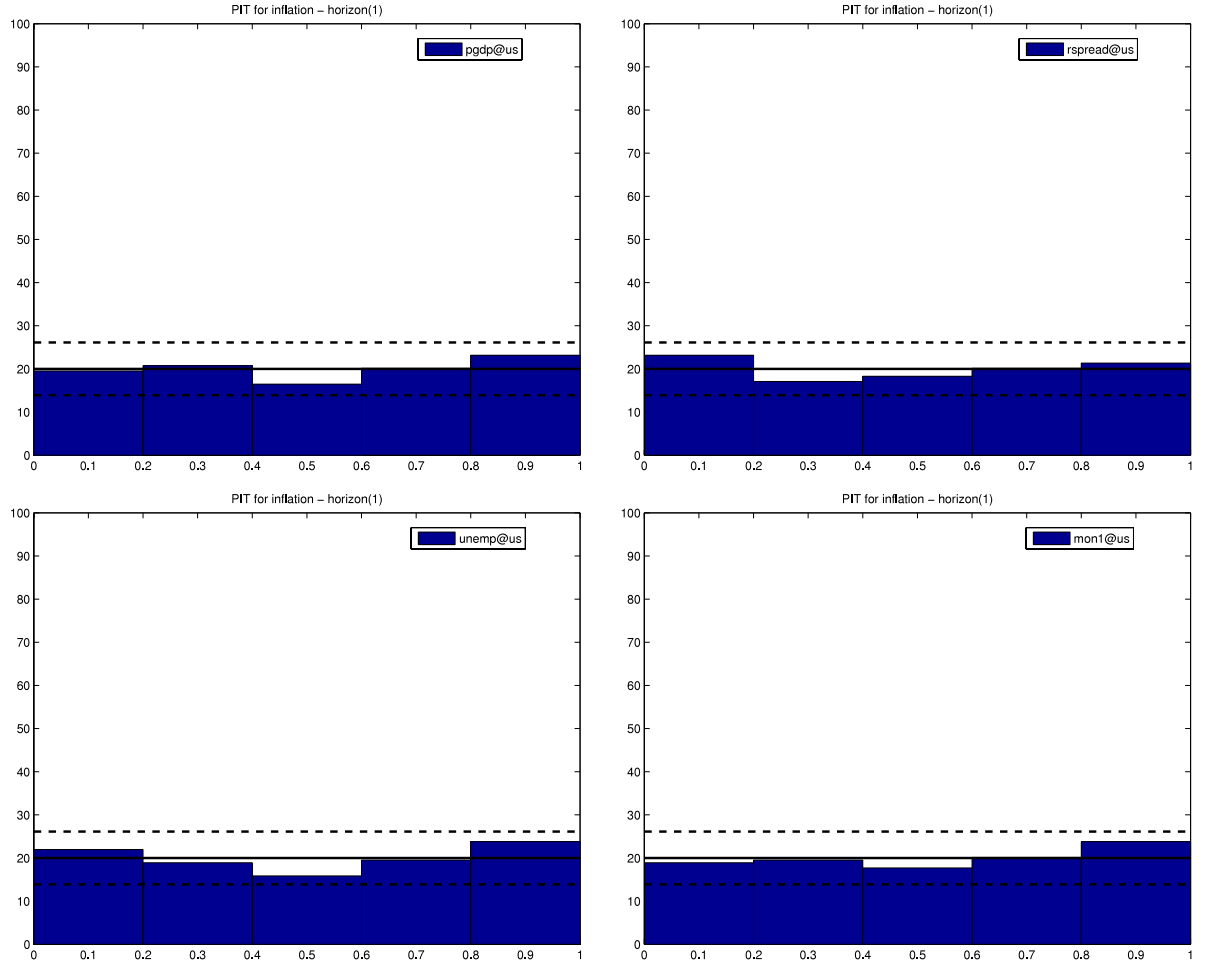
(b) Panel B: PITs for models combining large data sets for output growth at $h = 1$.

Fig. 1. (*continued*)

of 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 $\mathcal{Y}^h_{t+h} = [Y^h_{t+h,1},\ Y^h_{t+h,2},\ X^h_{t+h,1} \ldots X^h_{t+h,k} \ldots X^h_{t+h,K}]'$; $Y^h_{t+h,1}$ and $Y^h_{t+h,2}$ are the h-step-ahead variables for output growth and inflation, defined as in Eq. (8); $X^h_{k,t+h}$ =



(a) Panel A: PITs for ADL models of inflation at $h = 1$.

Fig. 2. Notes: The histograms depict the empirical distributions of the PITs. Solid lines represent the numbers of draws that are expected to be in each bin under a $U(0, 1)$ distribution. Dashed lines represent the 95% confidence intervals constructed under the normal approximation of a binomial distribution.

$(400/h)\sum_{j=1}^{h} X_{t+j}$; $y_t = [Y_{t,1}, Y_{t,2}, X_{t,1}, \ldots X_{t,k}, \ldots X_{t,K}]'$; $Y_{t,1}$ and $Y_{t,2}$ 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 $\Sigma_u = \text{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_{K+2}^2)$ over the respective rolling estimation windows.

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

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

and parameterize the prior such that it centers the regression coefficients around zero ($\bar{\beta} = 0$), reflecting our prior belief about the mean reverting nature of the variables in the VAR (all series except the rates are in first differences). Furthermore, $\bar{\Omega}$ 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:

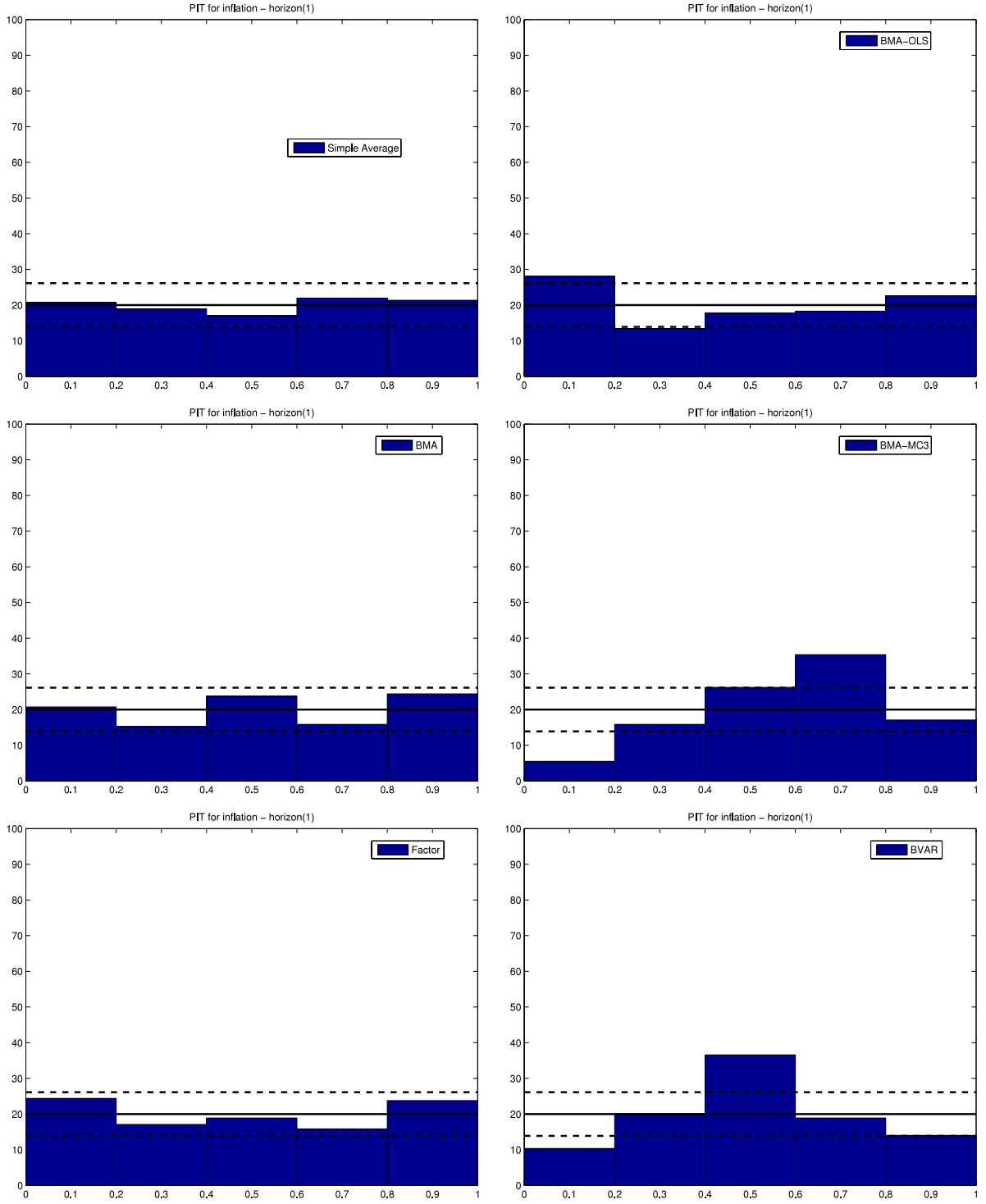
$$\text{Var}((B_l)_{ij}) = \begin{cases} \dfrac{\lambda^2}{l^2}, & j = i \\ \dfrac{\lambda^2}{l^2}\dfrac{\hat{\sigma}_i^2}{\hat{\sigma}_j^2}, & \text{otherwise,} \end{cases} \tag{26}$$

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 was recommended by Sims and Zha (1998). Given the results of 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 $i = 1, \ldots, K+2$ can be written as:

$$\Phi_{t+h}^{\text{BVAR}} = \Phi_{t+h}(y_{t+h,i}^h | y_t'\hat{B}_i, y_t'\hat{V}_i y_t + \hat{\sigma}_i), \tag{27}$$

where $\Phi_{t+h}(.)$ is the CDF of the normal distribution, $y = [1_{(K+2)}\ y_{l+1}\ \cdots\ y_T]$, $y_i^h = [y_{i,l+1}^h\ \cdots\ y_{i,T}^h]$, and



(b) Panel B: PITs for models combining large data sets for inflation at $h = 1$.

Fig. 2. (*continued*)

$$\hat{B}_i = (yy' + (\bar{\Omega}\lambda^2)^{-1})^{-1}(yy_i^{h'} + (\bar{\Omega}\lambda^2)^{-1}\bar{B}_i) \quad (28)$$

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

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

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 Avg.	0.37	0.35	Simple Avg.	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 the KS and AD tests following Kroese et al. (2011). *Indicates rejection at the 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 Avg.	0.15	0.36	Simple Avg.	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 per Kroese et al., 2011), based on four subsets $\{z_1, z_{1+h}, z_{1+2h}, \ldots\}$, $\{z_2, z_{2+h}, z_{2+2h}, \ldots\}$, $\{z_3, z_{3+h}, z_{3+2h}, \ldots\}$, and $\{z_4, z_{4+h}, z_{4+2h}, \ldots\}$. *Indicates rejection at the 5% significance level with Bonferroni bounds.

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 the Diebold et al. (1998), Kolmogorov–Smirnov, and Anderson–Darling tests. Next, we discuss tests of independence, and finally, we provide tests for identical distribution (instabilities). We conclude by considering tests for the correct specification, based on the inverse normal transformation of the PITs.

To preview our results, we find that there is more evidence against a lack of uniformity for density forecasts of inflation than for those of output growth, at both the 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, and in particular 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 the Bayesian model average, constructed using OLS estimates, perform fairly well in terms of the correct specification for output growth at the one-year-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, the forecast densities are generally 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 those of one-year-ahead ones. Several different versions of model averaging, as well as the factor model, perform fairly well, although 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.

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 Avg.	0.63	0.10	Simple Avg.	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*

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

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

Figs. 1 and 2 report results based on the Diebold et al. (1998) test for one-quarter-ahead density forecasts. Fig. 1 focuses on forecasts of output growth, whereas Fig. 2 considers inflation. In each figure, the graphs in Panel (A) report the empirical distributions of the PIT for ADL models, Eq. (8), with selected regressors (namely the lagged dependent variable, or autoregressive (AR) model; the spread; unemployment; and money (M1)). Panel (B), on the other hand, reports results for the PITs of various models when combining large data sets: the equal combinations of density forecasts across the ADL models, Eq. (9), labeled "Simple Average"; the BMA model with OLS parameter estimates, Eq. (10), labeled "BMA-OLS"; the BMA model with Bayesian parameter estimates, 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 graphs also report 95% confidence bands for the null hypothesis of i.i.d. uniformity.

The two panels of Fig. 1 show that, when forecasting output growth, Diebold et al.'s (1998) test rejects the hypothesis of normality (under the maintained assumption of independence) for the ADL model with the unemploy-

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

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 Avg.	0.46	0.36	Simple Avg.	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*

Note: The table reports minimum p-values of the LB test based on a $\chi^2(4)$ for four subsets $\{z_1, z_{1+h}, z_{1+2h}, \ldots\}$, $\{z_2, z_{2+h}, z_{2+2h}, \ldots\}$, $\{z_3, z_{3+h}, z_{3+2h}, \ldots\}$, and $\{z_4, z_{4+h}, z_{4+2h}, \ldots\}$.
*Indicates rejection at the 5% significance with Bonferroni bounds.

ment rate, the BMA model, BMA-MC3, and the BVAR. The histograms of the PITs of the BMA-MC3 and the BVAR suggest that more realizations fall in the middle of the distribution than would be expected if the PITs were i.i.d. uniform.

On the other hand, the two panels of Fig. 2 show 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)). The density combinations (reported in Panel (B)) appear to be calibrated well, with the exception of the BMA-OLS, BMA-MC3 and BVAR models: they again overestimate the realizations in the middle, but less severely than in the case of output growth.

For the same models, Tables 2 and 3 provide results for the Kolmogorov–Smirnov (labeled "KS") and 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 (particularly the AD test) also detect misspecification in all of the BMA model specifications, as well as the BVAR.[19] However, the simple average and the factor models are specified correctly. Note that the AD test is slightly more powerful in detecting misspecification in several models (e.g., the BMA-OLS and BMA), relative to the results reported in Fig. 1. The right panel in

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

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 Avg.	0.65		0.82		Simple Avg.	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 the 5% significance level.

Table 2 shows that, when predicting inflation, most of the predictors and models result in misspecified densities according to the AD test (although not according to the KS test); only the simple average and the BMA models are specified correctly according to both the KS and AD tests, while the densities of the factor and BMA-OLS models are misspecified according to the AD test. Note that the KS and AD tests often reach opposite conclusions: the discrepancies between the tests are most likely to be due to the higher power of the AD test relative to the KS test, especially in the tails of the distributions, to which we alluded in Section 2. The AD test finds more empirical evidence of misspecification in the case of inflation than for output growth. In addition, it also finds more evidence of misspecification than Diebold et al.'s (1998) test, especially for several ADL models. Overall, equally pooled models result in correctly specified densities according to all tests.

Table 3 shows the results for medium horizon (one-year-ahead) predictive densities. Due to the maintained assumption of independence and the serial correlation which is built into the four-quarter-ahead forecasts by construction, we divide the out-of-sample period into four subsets whose observations are four periods apart. For brevity, we only report the minimum p-values across the various subsets. The left panel shows that only ADL models which use exchange rates as predictors result in misspecified densities, and even then, 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 misspecification of the ADL models for inflation than for output growth at medium horizons: several nominal interest rate measures result in incorrectly specified densities. The tests reject the correct specification of several forecast combination models (including the BMS-OLS, BMA-MC3 and BVAR models), while they do not reject correct specification for the simple average and BMA models.

Overall, by comparing the right and left panels in the tables, under the maintained assumption of independence,

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 Avg.	0.22	0.08	Simple Avg.	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

Note: The table reports minimum p-values of the Andrews QLR test based on four subsets $\{z_1, z_{1+h}, z_{1+2h}, \ldots\}$, $\{z_2, z_{2+h}, z_{2+2h}, \ldots\}$, $\{z_3, z_{3+h}, z_{3+2h}, \ldots\}$, and $\{z_4, z_{4+h}, z_{4+2h}, \ldots\}$. *Indicates rejection at the 5% significance level with Bonferroni bounds.

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 the ADL models across Tables 2 and 3, we conclude that normality is more appropriate for forecasting output growth and inflation one year ahead than one quarter ahead. 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 except for forecasting inflation at the one-quarter-ahead horizon.

4.2. Tests of independence

The 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 of one quarter ($h = 1$). The left (right) panel in Table 4 reports results for forecasting output growth (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 (second column) and variance (third column) of $\{z_{t+h}\}_{t=R}^{T}$.

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 the ADL models with the T-bill rate). There is a significant level of serial correlation in the second moments of the PITs for the ADL model for several predictors (especially medium and long term 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 second moments of their PITs.

Turning to inflation, reported on in the right panel of Table 4, the most 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 of the density combinations (with the exception of BMA-OLS and BMA). In addition, there is no evidence of serial correlation in the first moments of the PITs for most of the 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.

Table 5 reports results for one-year-ahead density forecasts ($h = 4$). Due to the serial correlation which is built into the four-step-ahead forecasts by construction, we divide the out-of-sample period into four subsets with observations which are 4 periods apart. For the sake of brevity, we only 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 the test of no serial correlation in the PITs of inflation. For output growth, almost all of the ADL, simple average, BMA-OLS, BMA, factor, and 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 the inflation forecasts (especially in the second moments of the PITs) for the ADL models with employment, unemployment, and several money and interest rate measures. The simple average, BMA-MC3, factor and BVAR models also result in mis-calibrated densities. However, the BMA and BMA-OLS models do not show evidence of serial correlation in the PITs.

In general, forecast densities tend to be calibrated fairly well in terms of a lack of correlation 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 for one-year-ahead ones, as well as in second moments versus first. The most robust result in favor of correct specification across horizons and predictors again comes from the equal

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
rgdp@us	0.03*	0.55	0.06	pgdp@us	0.00*	0.94	0.00*
rovnght@us	0.00*	0.16	0.00*	rovnght@us	0.00*	0.80	0.00*
rtbill@us	0.00*	0.14	0.00 *	rtbill@us	0.00*	0.93	0.00*
rbnds@us	0.00*	0.17	0.00*	rbnds@us	0.00*	0.70	0.00*
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 Avg.	0.48	0.93	0.69	Simple Avg.	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 the Berkowitz LR test under various null hypotheses. *Indicates rejection at the 5% significance level.

averaging and BMA models, although simple averaging does show 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 Rossi (2013) and Stock and Watson (1996, 2003, 2007). Here, we are concerned that the distribution of the PITs might have changed over time. We therefore 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 the 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 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 with 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 also evidence of instability in the predictive densities of BMA, BMA-MC3, and BVAR models when predicting output growth. On the other hand, there is no evidence of instability in the predictive densities of the simple average, BMA-OLS and factor models. For the case of inflation, models based on pooling result in stable densities as well, with the

[20] As an alternative, one could also implement the BDS test of Broock, Scheinkman, and Dechert (1996). 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.

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 Avg.	0.02	0.50	0.04	Simple Avg.	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 $\{z_1, z_{1+h}, z_{1+2h}, \ldots\}$, $\{z_2, z_{2+h}, z_{2+2h}, \ldots\}$, $\{z_3, z_{3+h}, z_{3+2h}, \ldots\}$, and $\{z_4, z_{4+h}, z_{4+2h}, \ldots\}$. *Indicates rejection at the 5% significance level with Bonferroni bounds.

exception of the 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 $h = 4$.

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), not only can they test for uniformity and serial correlation jointly, they are also more powerful than the ones we reported earlier. Tables 8 and 9 report the results for Berkowitz's (2001) tests, while Table 10 reports results for the Doornik and Hansen (2008) test.

Interestingly, Tables 8 and 9 show that there is strong evidence of misspecification in the PITs for both output growth and inflation, at all horizons, according to Berkowitz's (2001) test for uniformity (labeled "$\mu = 0$, $\sigma = 1$"). Basically, the only models that are not misspecified for forecasting output growth at short horizons are the ADL model with exchange rate measures and the simple average, but 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 are the BMA, BMA-OLS, factor and simple average models. When predicting inflation one year ahead, only the ADL models based on nominal interest rates, and the BMA-OLS, BMA-MC3 and BVAR models, are not correctly specified.

We should note that Tables 8 and 9 also provide evidence of a lack of serial correlation in the PITs (columns 3 and 7, labeled "$\rho = 0$"), as well as evidence against the joint hypothesis of independence and normality of the inverse normal transform of the PITs (columns 4 and 8, labeled "joint"). The results for no serial correlation are in line with those implied by the Ljung–Box test, as reported in Tables 4 and 5. Serial correlation in

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 Avg.	0.03*	0.02	Simple Avg.	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*

Note: The table reports p-values of the Doornik–Hansen test for $h = 1$ and minimum p-values of the test for $h = 4$. *Indicates rejection at the 5% significance level (with Bonferroni bounds for the $h = 4$ case).

the first moment of the PITs is almost nonexistent for both the short and medium horizon predictive densities and 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 columns six and 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 any strong misspecification in the predictive densities of several ADL, BMA-OLS and factor models. However, based on this test, the simple average, BMA, BMA-MC3 and BVAR models appear to be misspecified. Most notably, the evidence of improper calibration is stronger for one-year-ahead density forecasts than for one-quarter-ahead ones.

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 each test, it summarizes the empirical evidence on the property listed in the corresponding column. For example, for the tests of uniformity listed in columns 2–5, "yes" indicates that uniformity is not rejected at the 5% significance level. The table shows that, for many models, the assumption of the normality of density forecasts is misspecified, according to at least one of the tests which we consider. The evidence in favor of the correct specification of normality is strongest for equally-weighted forecast averages, especially for predicting output growth at the one-year-ahead horizon, in which case none of the tests reject correct specification. The same is true for BMA for one-quarter-ahead inflation density forecasts. Overall, the performances of both models are more robust across target variables and horizons than those of all of the other models we consider. Thus, while each of the ADL models is misspecified 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 forecast combination models.

5. Conclusions

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

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", the 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 those of the authors solely, and should not be attributed to the Bank of Canada.

Table 11
Panel A. Summary of the results ($h = 1$).

	Uniformity		Un-correlation		Stability		Unif. and 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 Avg.	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 Avg.	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
Panel B. Summary of the results ($h = 4$).										
	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 Avg.	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 Avg.	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 or not 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 that 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 the 5% significance level, as listed in Tables 2–10.

References

Amisano, G., & Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business and Economic Statistics*, *25*(2), 177–190.

Anderson, T. W., & Darling, D. A. (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Annals of Mathematical Statistics*, *23*(2), 193–212.

Anderson, T. W., & Darling, D. A. (1954). A test of goodness-of-fit. *Journal of the American Statistical Association*, *49*(268), 765–769.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, *61*(4), 821–856.

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.

Bai, J., & Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business and Economic Statistics*, *25*(1), 52–60.

Banerjee, A., & Marcellino, M. (2006). Are there any reliable leading indicators for the US inflation and GDP growth? *International Journal of Forecasting*, *22*, 137–151.

Banerjee, A., Marcellino, M., & Masten, I. (2005). Leading indicators for euro-area inflation and GDP growth. *Oxford Bulletin of Economics and Statistics*, *67*(S1), 785–813.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, *19*(4), 465–474.

Broock, W. A., Scheinkman, J. A., & Dechert, W. D. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, *15*(3), 197–235.

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.

Clements, M. P., & Smith, J. (2000). Evaluating the forecast densities of linear and non-linear models: applications to output growth and unemployment. *Journal of Forecasting*, *19*(4), 255–276.

Corradi, V., & Swanson, N. R. (2006). Predictive density evaluation. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting, vol. 1* (pp. 197–284). North Holland: Elsevier.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, *39*(4), 863–883.

Diebold, F. X., Tay, A. S., & Wallis, K. F. (1999). Evaluating density forecasts of inflation: the Survey of Professional Forecasters. In R. F. Engle, & H. White (Eds.), *Cointegration, causality, and forecasting: a festschrift in honour of Clive W.J. Granger* (pp. 76–90). Oxford University Press.

Diks, C., Panchenko, V., & van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, *163*(2), 215–230.

Doornik, J. A., & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, *70*(S1), 927–939.

Garratt, A., Lee, K., Pesaran, M. H., & Shin, Y. (2003). Forecast uncertainties in macroeconomic modeling: an application to the UK economy. *Journal of the American Statistical Association*, *98*(464), 829–838.
Giannone, D., Lenza, M., & Primiceri, G. (2012). *Prior selection for vector autoregressions*. Mimeo.
Granger, C., & Pesaran, M. H. (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting*, *19*(7), 537–560.
Guidolin, M., & Timmermann, A. (2006). Term structure of risk under alternative econometric specifications. *Journal of Econometrics*, *131*(1–2), 285–308.
Hayashi, F. (2000). *Econometrics*. Princeton University Press.
Jore, A. S., Mitchell, J., & Vahey, S. P. (2010). Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics*, *25*(4), 621–634.
Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. In *Giornale dell'Istituto Italiano Degli Attuari, vol. 4* (pp. 83–91).
Koop, G. (2003). *Bayesian econometrics*. John Wiley.
Kroese, D. P., Taimre, T., & Botev, Z. I. (2011). *Handbook of Monte Carlo methods*. John Wiley.
Manzan, S., & Zerom, D. (2013). Are macroeconomic variables useful for forecasting the distribution of U.S. inflation? *International Journal of Forecasting*, *29*(3), 469–478.
Marcellino, M., Stock, J. H., & Watson, M. W. (2003). Macroeconomic forecasting in the Euro area: country specific versus Euro wide information. *European Economic Review*, *47*(1), 1–18.
Mitchell, J., & Wallis, K. F. (2011). Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics*, *26*(6), 1023–1040.
Newey, W. K., & West, K. O. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708.
Noceti, P., Smith, J., & Hodges, S. (2003). An evaluation of tests of distributional forecasts. *Journal of Forecasting*, *22*(6–7), 447–455.
Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, *23*(3), 470–472.
Rossi, B. (2013). Advances in forecasting under instabilities. In G. Elliott, & A. Timmermann (Eds.), *Handbook of economic forecasting, vol. 2*. Elsevier-North Holland Publications.
Rossi, B., & Sekhposyan, T. (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.
Rossi, B., & Sekhposyan, T. (2013). Conditional predictive density evaluation in the presence of instabilities. *Journal of Econometrics*, forthcoming.
Sims, C. A., & Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, *39*(4), 949–968.
Smirnov, N. V. (1948). Tables for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, *19*(2), 279–281.
Stock, J. H., & Watson, M. W. (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics*, *14*(1), 11–30.
Stock, J. H., & Watson, M. W. (2002). *Introduction to econometrics*. Addison-Wesley.
Stock, J. H., & Watson, M. W. (2003). Forecasting output and inflation: the role of asset prices. *Journal of Economic Literature*, *41*(3), 788–829.
Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven country data set. *Journal of Forecasting*, *23*(6), 405–430.
Stock, J. H., & Watson, M. W. (2007). Has inflation become harder to forecast? *Journal of Money, Credit and Banking*, *39*(1), 3–33.
Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting, vol. 1* (pp. 135–196). North Holland: Elsevier.
Wright, J. H. (2009). Forecasting US inflation by Bayesian model averaging. *Journal of Forecasting*, *28*(2), 131–144.

Barbara Rossi is an ICREA professor at UPF and Barcelona GSE, and associate researcher at CREI. She specializes in the fields of time series econometrics, as well as applied international finance and macroeconomics. Her current research focuses on forecasting and macroeconometrics. Prof. Rossi has published her research findings in the *Review of Economic Studies*, *QJE*, the *International Journal of Forecasting*, *JBES*, the *International Economic Review*, *Econometric Theory*, the *Journal of Applied Econometrics*, *JMCB*, *Journal of Econometrics*, and the *Review of Economics and Statistics*, among others. She holds associate editorial positions at *JBES*, *JAE* and *JEDC*.

Tatevik Sekhposyan is a Senior Analyst at the Bank of Canada. Her main research interests are Macroeconomics, Monetary Theory, Time Series and Bayesian Econometrics. Dr. Sekhposyan has held visiting positions at the St. Louis Fed and Atlanta Fed. She has published her works in the *B.E. Journal of Macroeconomics*, *Journal of Econometrics* and the *International Journal of Forecasting*. Dr. Sekhposyan has presented her works at the Society for Nonlinear Dynamics and Econometrics Annual Symposium, the AEA Meetings, the Conference on Computational and Financial Econometrics, and the (EC)2, among other conferences.