Alternative Tests for Correct Specification of Conditional Predictive Densities

Barbara Rossi[1] and Tatevik Sekhposyan[2]

March 28, 2015

Abstract

We propose new methods for evaluating predictive densities in an environment where the estimation error of the parameters used to construct the densities is preserved asymptotically under the null hypothesis. The tests offer a simple way to evaluate the correct specification of predictive densities. Monte Carlo simulation results indicate that our tests are well sized and have good power in detecting misspecifications. An empirical application to the Survey of Professional Forecasters and a baseline macroeconomic model shows the usefulness of our methodology.

Keywords: Predictive Density, Dynamic Misspecification, Forecast Evaluation
J.E.L. Codes: C22, C52, C53

Acknowledgments: We thank T. Clark, F. Diebold, G. Ganics, A. Inoue, A. Patton, B. Perron, F. Ravazzolo, N. Swanson, M. Watson, seminar participants at George Washington Univ., Lehigh Univ., Univ. of California-Riverside, University of Mississippi, Texas A&M, Rutgers, CORE Louvain-la-Neuve, Univ. of New South Wales, Monash, ECB and the 2012 Econometric Society Australasian Meetings, the 2012 Time Series Econometrics Workshop in Zaragoza, the 2013 CIREQ Time Series and Financial Econometrics Conference, the 2013 St. Louis Fed Applied Time Series Econometrics Workshop, the 2013 UCL Conference on Frontiers in Macroeconometrics, the 2013 Conference on Forecasting Structure and Time Varying Parameter Patterns, the 2014 EC^2 Conference and the 2014 CAMP Workshop on Empirical Macroeconomics for comments. B. Rossi gratefully acknowledges financial support from the European Research Agency's Marie Curie Grant 303434 and ERC Grant 615608.

[1]ICREA-University of Pompeu Fabra, Barcelona GSE and CREI, c/Ramon Trias Fargas 25/27, Barcelona 08005, Spain; tel.: +34-93-542-1655; e-mail: barbara.rossi@upf.edu

[2]Texas A&M University, 3060 Allen Building, 4228 TAMU, College Station, TX 77843, USA; tel.: +1-979-862-8857; e-mail: tsekhposyan@tamu.edu

1 Introduction

Policy institutions are becoming interested in complementing point forecasts with an accurate description of uncertainty. For instance, they are interested not only in knowing whether inflation is below its target, but also in understanding whether the realized inflation rate was forecasted to be a low probability event ex-ante. In fact, if researchers underestimate the uncertainty around point forecasts, it is possible that an event with a fairly high likelihood of occurrence is forecasted to be a very low probability event. An accurate description of uncertainty is therefore important in the decision making process of economic agents and policymakers. The interest in density forecasting has emerged in the survey by Elliott and Timmermann (2008) as well as in their recent book (Elliott and Timmermann, 2014), and has inspired several empirical contributions that have proposed new approaches to improve the forecasting performance of predictive densities, e.g. Ravazzolo and Vahey (2014), Aastveit, Foroni and Ravazzolo (2014), and Billio, Casarin, Ravazzolo and van Dijk (2013). The objective of this paper is to provide reliable tools for evaluating whether the uncertainty around point forecasts, and predictive densities in general, are correctly specified.

Many central banks periodically report fan charts to evaluate and communicate the uncertainty around point forecasts (e.g., see the various issues of Bank of England Inflation Report or the Economic Bulletin of the Bank of Italy[3]). Fan charts provide percentiles of the forecast distribution for variables of interest. Typically, central banks' fan charts are the result of convoluted methodologies that involve a variety of models and subjective assessments, although fan charts can be based on specific models as well.[4]

INSERT FIGURE 1 HERE

Figure 1 plots fan charts for US output growth (left panel) and the Federal Funds rate (right panel) based on a representative macroeconomic model widely used in academia and policymaking (discussed in detail later on). The fan charts display model-based forecasts made in 2000:IV for the next four quarters. The shaded areas in the figures depict the deciles of the forecast distribution and provide a visual impression of the uncertainty around the point forecasts (in this case, the median, marked by a dashed line). Over the four quarterly horizons, uncertainty about output growth and interest rate forecasts has a very different

[3] These publications are available at http://www.bankofengland.co.uk/publications/Pages/inflationreport and https://www.bancaditalia.it/pubblicazioni/econo/bollec, respectively.

[4] See for instance Clements (2004) for a discussion on the Bank of England fan charts.

pattern: the uncertainty surrounding output growth forecasts is constant across horizons, while for interest rates it depends on the horizon. The dark, solid line in the figures plots the actual realized value of the target variable. Clearly, forecasts of interest rates were very imprecise (the realization is outside every forecast decile except for one-quarter-ahead horizon), whereas the model predicts output growth more accurately. In order to evaluate the model-based forecast distributions, it is important to understand whether it is the description of uncertainty that was inaccurate or the realized values were indeed low probability events.

Currently available methodologies test whether the empirical distribution belongs to a given parametric density family with parameters evaluated at their pseudo-true values. Our paper derives new tools to evaluate whether predictive densities are correctly specified by focusing on evaluating their actual forecasting ability conditional on models' estimated parameter values. In other words, we test whether the predictive densities are correctly specified given the considered parametric model and its estimation technique. Accordingly, our test does not require an asymptotic correction for parameter estimation error. Importantly, our tests do not require the model to be dynamically correctly specified nor its disturbances to be serially uncorrelated; however, for completeness, we also discuss a version of the tests that hold when the model is dynamically correctly specified.

The advantage of this approach relative to the existing literature is that it allows the researcher to evaluate whether the density forecast is correctly specified conditional on the actual parameter estimates. In contrast, most of the literature focuses on testing the correct specification of predictive densities evaluated at the pseudo-true parameter values, which may not be representative of the models' actual forecasting ability in finite samples. We propose an approach where parameter estimation error is maintained under the null hypothesis, as in Amisano and Giacomini (2007). However, our approach is very different, as the latter focus on model selection by comparing the *relative performance* of competing models' predictive densities, whereas we focus on evaluating the *absolute performance* of a model's predictive density.

Maintaining parameter estimation error under the null hypothesis has two advantages: (i) there is no need to correct the asymptotic distribution of test statistics for parameter estimation error, since that is maintained under the null hypothesis; and (ii) the asymptotic distribution of the test statistics at the one-step-ahead horizon is nuisance parameter free and the critical values can be tabulated when the model is dynamically correctly specified.[5] We derive our tests within a class of Kolmogorov-Smirnov and Cramér-von Mises-type

[5]Note that (ii) is not unique to cases where parameter estimation error is maintained under the null

tests commonly used in the literature and show that all our proposed tests have good size properties in small samples.

When misspecification of the predictive density is detected, an important step is to understand the source of the misspecification. Many tests that exist in the literature concentrate on testing for correct specification of predictive densities by testing a joint hypothesis of uniformity and independence. Lack of uniformity implies an incorrect unconditional probability (on average) that the actual realizations of the target variable match the model's predictive density. Lack of independence refers to a situation where, even if on average realizations are compatible with the model's predictive density (i.e., the unconditional probability is correct), the pattern of the rejections is non-random. Thus, the rejection of the correct specification could be driven either by the lack of uniformity or independence, and it is important to identify the source. To uncover the source of the misspecification, we: (i) propose new tests of uniformity robust to violations of independence; (ii) discuss some tests of serial correlation robust to violations of uniformity that are available in the literature and could be used. The tests can be applied to either one-step-ahead or multiple-step-ahead predictive densities.

Our paper is related to a series of contributions which test whether observed forecasts could have been generated by a given distribution. Diebold et al. (1998, 1999) introduced the probability integral transform (PIT) into economics as a tool to test whether the empirical predictive distribution of surveys or empirical models matches the true, unobserved distribution that generates the data. Their approach tests for properties of the PITs, such as independence and uniformity, by treating the forecasts as primitive data, that is without correcting for estimation uncertainty associated with those forecasts. Additional approaches proposed in the literature for assessing the correct calibration of predictive densities are the non-parametric approach by Hong and Li (2005) and the bootstrap introduced by Corradi and Swanson (2006 a,b,c).[6] The null hypothesis in the latter is that of correct specification of the density forecast at the pseudo-true (limiting) parameter values. Although this framework enables predictive density evaluation when the models are dynamically misspecified, it does not necessarily capture the actual measure of predictive ability that researchers are interested in, as in small samples the pseudo-true parameter values may not be representative of the actual marginal predictive ability of the regressors. In the approach we propose, the

hypothesis; in fact, it also holds when parameter estimation error is asymptotically irrelevant, or when one uses martingalization techniques, as in Bai (2003).

[6] Hong, Li and Zhao (2007) provide with an out-of-sample counterpart of the Hong and Li (2005) in-sample tests. See also Bontemps and Meddahi (2012) for in-sample tests of distributional assumptions.

main test statistic is the same as Corradi and Swanson's (2006a) one, although the null hypothesis is very different: it targets evaluating density forecasts at the estimated parameter values (as opposed to their population values). A more recent alternative has been proposed by González-Rivera and Sun (2014); they use graphical devices to implement a test of correct specification. The proposed methods work when models are dynamically correctly specified, however, when parameter estimation error is asymptotically relevant, the asymptotic distribution is not nuisance parameter free and a bootstrap procedure is proposed. Our test, instead, does not require a bootstrap procedure, and its critical values are readily available in the case when the models are dynamically correctly specified.[7]

We provide empirical applications of our proposed tests to the density forecasts in the Survey of Professional Forecasters (SPF) as well as those produced by a baseline DSGE model. We find that predictive densities are, in general, misspecified.

The remainder of the paper is organized as follows. Section 2 introduces the notation and definitions. Section 3 presents results for tests of correct specification of density forecasts robust to dynamic misspecification and Section 4 discusses issues related to the practical applicability of our test. In Section 5, we provide Monte Carlo evidence on the performance of our tests in small samples. Section 6 analyzes the empirical applications to SPF and DSGE density forecasts and Section 7 concludes.

2 Notation and Definitions

We first introduce the notation and discuss the assumptions about the data, the models and the estimation procedure. Consider a stochastic process $\{Z_t : \Omega \to R^{k+1}\}_{t=1}^{T}$ defined on a complete probability space (Ω, F, P). The observed vector Z_t is partitioned as $Z_t = (y_t, X_t')'$, where $y_t : \Omega \to R$ is the variable of interest and $X_t : \Omega \to R^k$ is a vector of predictors. Let $1 \leq h < \infty$. We are interested in the true but unknown $h-$step-ahead conditional predictive density for the scalar variable y_{t+h} based on $F_t = \sigma(Z_1', ..., Z_t')'$, which is the true information set available at time t. We denote this density by $\phi_0(.)$.[8]

[7] We should note that allowing for dynamic misspecification under the null makes the test robust to violations of independence, which is important since the test can then be used for evaluating multi-step ahead densities. This is different from the large number of tests suggested in the literature which are not applicable to test the correct calibration of multi-step-ahead densities (e.g. Diebold et al., 1998, and Gonzalez-Rivera and Sun, 2014).

[8] The true conditional forecast density may depend on the forecast horizon. To simplify notation, we omit this dependence without loss of generality given that the forecast horizon is fixed. Furthermore, we use the

We assume that the researcher has divided the available sample of size $T+h$ 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 out-of-sample density forecasts of the variable of interest y_t using the information set $\Im_t$, such that $R+P-1+h=T+h$ and $\Im_t \subseteq \mathcal{F}_t$. Note that this implies that the researcher observes a subset of the true information set. We also let $\Im_{t-R+1}^t$ denote the truncated information set between time $(t-R+1)$ and time t used by the researcher.

Let the sequence of P out-of-sample estimates of conditional predictive densities evaluated at the ex-post realizations be denoted by $\left\{\phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^t\right)\right\}_{t=R}^T$. The dependence on the information set is a result of the assumptions we impose on the in-sample parameter estimates, $\widehat{\theta}_{t,R}$. We assume that the parameters are re-estimated at each $t=R,...,T$ over a window of R data including data indexed $t-R+1,...,t$ (rolling scheme).[9] In this paper we are concerned with direct multi-step forecasting, where the predictors are lagged h periods. In addition to being parametric (such as a normal distribution), the distribution $\phi_{t+h}(.)$ can also be non-parametric (as in one of the empirical applications in this paper).

Consider the probability integral transform (PIT), which is the cumulative density function (CDF) corresponding to $\phi_{t+h}(.)$ evaluated at the realized value y_{t+h}:

$$z_{t+h}=\int_{-\infty}^{y_{t+h}}\phi_{t+h}\left(y|\Im_{t-R+1}^t\right)dy\equiv\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^t\right).$$

Let

$$\xi_{t+h}(r)\equiv\left(1\left\{\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^t\right)\leq r\right\}-r\right),$$

where $1\{.\}$ is the indicator function and $r\in[0,1]$. Consider $\Psi(r)=\Pr\{z_{t+h}\leq r\}-r$ and its out-of-sample counterpart:

$$\Psi_P(r)\equiv P^{-1/2}\sum_{t=R}^{T}\xi_{t+h}(r). \tag{1}$$

Let us also denote the empirical probability distribution function of the PIT by

$$\varphi_P(r)\equiv P^{-1}\sum_{t=R}^{T}1\left\{\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^t\right)\leq r\right\}. \tag{2}$$

symbols $\phi_0(.)$ and $\phi_t(.)$ to denote generic distributions and not necessarily a normal distribution.

[9] The choice of the estimation scheme (rolling versus recursive) depends on the features of the data: in the presence of breaks, one would favor a rolling scheme that allows a fast update of the parameter estimates, at the cost of a potential increase in estimation uncertainty relative to a recursive scheme when there are not break. As discussed in Giacomini and White (2006), our proposed approach is also valid for other classes of limited memory estimators.

3 Asymptotic Tests of Specification

This section presents results for the case of one-step-ahead forecasts when the densities are dynamically correctly specified; we then generalize the tests to the presence of misspecification and serial correlation. The generalized case could also apply to the $h > 1$ step-ahead forecasts. All the proofs are relegated to Appendix A. The tests we propose have an asymptotic distribution that is free of nuisance parameters in the one-step-ahead forecast case when the models are dynamically correctly specified. In this case the critical values can be tabulated. We also discuss tests that are valid for multi-step-ahead forecasts and in the presence of dynamic misspecification. Both of these cases introduce serial correlation in the dynamics of the PITs.

In order to maintain parameter estimation error under the null hypothesis, we state our null hypothesis in terms of a truncated information set, which expresses the dependence of the predictive density on estimated parameter values (as in Amisano and Giacomini, 2007). We focus on testing $\phi_{t+h}\left(y|\Im_{t-R+1}^{t}\right) = \phi_0\left(y|\mathcal{F}_t\right)$, that is:

$$H_0 : \Phi_{t+h}\left(y|\Im_{t-R+1}^{t}\right) = \Phi_0\left(y|\mathcal{F}_t\right) \text{ for all } t = R, ..., T, \tag{3}$$

where $\Phi_0\left(y|\mathcal{F}_t\right) \equiv \Pr\left(y_{t+h} \leq y|\mathcal{F}_t\right)$ denotes the distribution specified under the null hypothesis.[10] The alternative hypothesis, H_A, is the negation of H_0. Note that the null hypothesis evaluates the correct specification of the density forecast of a model estimated with a given window size, R, as well as the parameter estimation method chosen by the researcher.

We are interested in the test statistics:

$$\kappa_P = \sup_{r\in[0,1]} \Psi_P\left(r\right)^2, \tag{4}$$

$$C_P = \int_0^1 \Psi_P\left(r\right)^2 dr. \tag{5}$$

Note that the κ_P test statistic is basically the same as the V_{1T} test statistic considered by Corradi and Swanson (2006a) when applied to predictive densities (the latter consider the absolute value of $\Psi_P\left(r\right)$, while we consider its square). Note, however, that we derive

[10] Note that the null hypothesis depends on R. In other words, the null hypothesis jointly tests density functional form and estimation technique. It might be possible that correct specification is rejected for a model for some values of R and not rejected for the same model for some other choices of R. This is reasonable since we are evaluating the model's performance when estimated in a given sample size, so the estimation error is important under the null hypothesis. Alternatively, one could construct a test that is robust to the choice of the estimation window size as suggested in Inoue and Rossi (2012) and references therein.

the asymptotic distribution of the test statistic under a different null hypothesis. Corradi and Swanson (2006a) focus on the null hypothesis: $H_0^{CS} : \Phi_{t+h}(y|\Im_t) = \Phi_0(y|\Im_t, \theta^\dagger)$ for some $\theta^\dagger \in \Theta$, where Θ is the parameter space. That is, the latter test the hypothesis of correct specification of the predictive density at the pseudo-true parameter value. Thus, the limiting distribution of their test reflects parameter estimation error and, therefore, is not nuisance parameter free. In addition, they allow for dynamic misspecification under the null hypothesis. This allows them to obtain asymptotically valid critical values even when the information set may not contain all the relevant past history. Dynamic misspecification also affects the limiting distribution of their test statistic by contributing additional nuisance parameters.

Under our null hypothesis (eq. 3) instead, the limiting distribution of the test statistic is nuisance parameter free when the model is dynamically correctly specified. The reason is that we maintain parameter estimation error under the null hypothesis, which implies that the asymptotic distribution of the test does not require a delta-method approximation around the pseudo-true parameter value.

To clarify our null hypothesis, we provide an example.

Example: As a simple example, consider $y_{t+1} = c_t + x_t + \varepsilon_{t+1}$, $\varepsilon_{t+1} \sim iid\ N(0,1)$ and $x_t \sim iid\ N(0, \sigma_x^2)$, and ε_{t+1}, x_t are independent of each other. We assume for simplicity that the variance of the errors is known and equals one. The researcher instead considers a model $y_{t+1} = \beta x_t + e_{t+1}$, $e_{t+1} \sim iid\ N(0,1)$. Moreover, the researcher is estimating the coefficient β with a window of size R. We set c_t such that our null hypothesis (eq. 3) holds. That is, the estimated PIT is:

$$\int_{-\infty}^{y_{t+1}} \phi_{t+1}\left(y|\Im_{t-R+1}^{t}\right) dy,$$

where $\phi_{t+1}\left(y|\Im_{t-R+1}^{t}\right)$ is $N\left(\widehat{\beta}_{t,R}x_t, 1\right)$, whereas the PIT that generated the data is:

$$\int_{-\infty}^{y_{t+1}} \phi_{t+1}\left(y|\mathcal{F}_t\right) dy,$$

where $\phi_{t+1}\left(y|\mathcal{F}_t\right)$ is $N\left(c_t + x_t, 1\right)$. Under the assumption that the variance is known, a sufficient condition for the null hypothesis to hold is that the conditional means from true DGP and the estimated model are the same. More in detail, the null hypothesis is imposed by assuming:

$$c_t + x_t = \widehat{\beta}_{t,R}x_t,$$

that is,[11]

$$c_t = \left(\frac{R^{-1} \sum_{j=t-R+1}^{t} \left[x_{j-1} - \left(R^{-1} \sum_{s=t-R+1}^{t} x_{s-1} \right) \right] \left[y_j - \left(R^{-1} \sum_{s=t-R+1}^{t} y_s \right) \right]}{R^{-1} \sum_{j=t-R+1}^{t} \left[x_{j-1} - \left(R^{-1} \sum_{s=t-R+1}^{t} x_{s-1} \right) \right]^2} - 1 \right) x_t.$$

Thus, the null hypothesis in eq. (3) does not test the correct specification of the forecasting model evaluated at the true parameter values (relative to the data generating process); rather, the null hypothesis in eq. (3) tests the correct specification of the forecasting model evaluated at the parameter values obtained conditional on the estimation procedure. We argue that the latter is an appropriate approach to evaluate the correct specification of density forecasts, since it jointly evaluates the proposed model and its estimation technique, including the estimation window size. The methodology only requires that the conditional mean be estimated based on a finite number of observations.[12]

Suppose, instead, the true data generating process is: $y_t = \alpha + x_t + \varepsilon_t$ where $x_t \sim iid\chi_1^2$ and $\varepsilon_t \sim iidN(0,1)$. Let the researcher estimate the a misspecified model that includes only a constant treating the forecast distribution as normal. Note that the null hypothesis does not hold even if the error term is normal, since the misspecification results in an actual error term that is a combination of x_t and ε_t. Thus, since the data is generated as a mixture of a chi-squared and normal distribution, and we are testing whether it is a normal, the null hypothesis does not hold.

3.1 One-step-ahead Density Forecasts and Dynamically Correctly Specified Models

This sub-section presents results for the case of one-step-ahead forecasts when the densities are dynamically correctly specified; the next sub-section generalizes the tests to the presence of misspecification and serial correlation. Let $h = 1$. First, we derive the asymptotic distribution of $\Psi_P(r)$ for one-step-ahead density forecasts under Assumption 1.[13]

[11] The data under the null hypothesis are mixing, and thus satisfy our Assumption 1, for the following reason: let $g_t \equiv (x_t, c_t, \varepsilon_t)'$. Since $E(g_t) = 0$ and $E(g_t|g_{t-1}, g_{t-2}, ...) = 0$ then g_t is a martingale difference sequence and has finite variance, thus it is white noise (Hayashi, p. 104).

[12] The results in this paper also carry over to the fixed-estimation scheme, where the conditioning information set is $\Im_1^R$, or to any other information set based on a bounded number of observations R, provided R is finite.

[13] Note that if P/R$\to$ 0, our test would be the same as the existing tests as parameter estimation uncertainty becomes irrelevant in those cases (see Corradi and Swanson, 2006b). This result would hold even for recursive estimation schemes as long as P/R$\to$ 0. However, we test a different null hypothesis than the existing tests.

Assumption 1.

(i) $\left\{Z_t = (y_t, X_t')'\right\}_{t=R}^{T}$ is strong mixing with mixing coefficients $\alpha(m)$ of size $-\lambda/(\lambda-1)$, where $\lambda \in (1, 3/2)$;

(ii) $\Phi_0(y_{t+h}|\mathcal{F}_t)$ is continuous, differentiable and has a well defined inverse;

(iii) $F_d(.,.)$ and $F(.)$ are respectively the joint and the marginal distribution functions of the random variable $\Phi_0(y_{t+h}|\mathcal{F}_t)$, i.e. $\Pr(\Phi_0(y_{t+h}|\mathcal{F}_t) \leq r_1, \Phi_0(y_{t+h+d}|\mathcal{F}_{t+d}) \leq r_2) = F_d(r_1, r_2)$, $\Pr(\Phi_0(y_{t+h}|\mathcal{F}_t) \leq r) = F(r)$, and $F(r)$ is continuous;

(iv) $R < \infty$ as $P, T \to \infty$.

Assumption 1(i) allows for short memory heterogeneous data. The assumption is similar to that in Theorem 1 in Giacomini and White (2006), which allows for some types of mild non-stationarity induced by changes in distributions over time, yet rules out I(1) processes, a useful extension to the existing literature, since the parametric tests of correct specification that allow for dynamic misspecification under the null assume covariance stationary data. Assumption 1(ii) and 1(iii) are similar to those maintained in Inoue (2001), Assumption B. These assumptions require the PITs, as well as the marginal and joint distributions of the PITs, are well-defined.[14] Assumption 3(iv) assumes the estimation window size stays finite as the total sample size grows. Note that the assumption potentially allows forecasts to be conditioned on a finite set of future values of some variables of interest (i.e. "conditional forecasts").

Dynamic correct specification is characterized by Assumption 2(a):

Assumption 2.

(a) $y_{t+h}|\Im_{t-R+1}^{t} \equiv y_{t+h}|\mathcal{F}_t$ *for all* $t = R, ..., T$, *where* $\equiv$ *denotes equality in distribution; and (b)* $\Phi_{t+1}^{-1}\left\{z_{t+1}|\Im_{t-R+1}^{t}\right\}_{t=R}^{T}$ *has non-zero Jacobian with continuous partial derivatives.*

We show the following result:

Theorem 1 (Asymptotic Distribution of $\Psi_P(r)$) *Under Assumptions 1, 2, and* H_0 *in eq. (3): (i)* $\{z_{t+1}\}_{t=R}^{T}$ *is iid* $U(0,1)$*; (ii)* $\Psi_P(r)$ *weakly converges as a variable in the space* $([0,1] \times \mathbb{R})$ *to the Gaussian process* $\Psi(.)$*, with mean zero and auto-covariance function* $E[\Psi(r_1)\Psi(r_2)] = [\inf(r_1, r_2) - r_1 r_2]$.

[14] The assumption is on the unobserved true distribution, though under the null it also ensures that the proposed distribution has a well defined limiting distribution.

The result in Theorem 1 allows us to derive the asymptotic distribution of the test statistics of interest, presented in Theorem 2. The latter shows that the asymptotic distribution of our proposed test statistics have the appealing feature of being nuisance parameter free.

Theorem 2 (Correct Specification Tests) *Under Assumptions 1, 2 and H_0 in eq. (3):*

$$\kappa_P \equiv \sup_{r\in[0,1]} \Psi_P(r)^2 \Rightarrow \sup_{r\in[0,1]} \Psi(r)^2, \tag{6}$$

and

$$C_P \equiv \int_0^1 \Psi_P(r)^2 dr \Rightarrow \int_0^1 \Psi(r)^2 dr. \tag{7}$$

The tests reject H_0 at the $\alpha\cdot 100\%$ significance level if $\kappa_P > \kappa_\alpha$ and $C_P > C_\alpha$. Critical values for $\alpha = 10\%$, 5% and 1% are provided in Table 1, Panel A.

INSERT TABLE 1 HERE

Note that one could be interested in testing correct specification in specific parts of the distribution.[15] For example, one might be interested in the tails of the distribution, which correspond to outliers, such as the left tail where $r \in [0, 0.25)$, or the right tail where $r \in [0.75, 1)$, or both: $r \in \{[0, 0.25 \cup 0.75, 1]\}$. Alternatively, one might be interested in the central part of the distribution, for example $r \in [0.25, 0.75]$. We provide critical values for these interesting cases in Table 1, Panel B.

Note also that our κ_P test has a graphical interpretation. In fact,

$$\alpha = \Pr\left\{\sup_{r\in[0,1]} \Psi_P(r)^2 > \kappa_\alpha\right\} = \Pr\left\{\left[\sup_{r\in[0,1]} |\Psi_P(r)|\right]^2 > \kappa_\alpha\right\} = \Pr\left\{\sup_{r\in[0,1]} |\Psi_P(r)| > \sqrt{\kappa_\alpha}\right\}.$$

Thus, from eqs. (1) and (2),

$$\frac{1}{\sqrt{P}}\Psi_P(r) \equiv P^{-1}\sum_{t=R}^{T}\left(1\left\{\Phi_{t+h}\left(y_{t+h}|\Im_{t-R+1}^{t}\right) \le r\right\} - r\right) = \varphi_P(r) - r.$$

Furthermore,

$$\alpha = \Pr\left\{\sup_{r\in[0,1]} |\Psi_P(r)| > \sqrt{\kappa_\alpha}\right\} = \Pr\left\{\sup_{r\in[0,1]} |\varphi_P(r) - r| > \sqrt{\kappa_\alpha/P}\right\}.$$

[15] See Franses and van Dijk (2003), Amisano and Giacomini (2007) and Diks, Panchenkob and van Dijk (2011) for a similar idea in the context of point forecasts and density forecast comparisons.

This suggests the following implementation: plot the cumulative distribution function of the PIT, eq. (2), together with the cumulative distribution function of the uniform, r (the 45-degree line), and the critical value lines: $r \pm \sqrt{\kappa_\alpha / P}$. Then, the κ_P test rejects if the cumulative distribution function of the PIT is outside the critical value lines. It also follows from this argument that the critical values of the test statistic $\sup_{r \in [0,1]} |\Psi_P(r)|$ would be $\sqrt{\kappa_\alpha}$.

It is interesting to compare our approach to Diebold et al. (1998). While our null hypothesis is different from theirs, the procedure that we end up proposing is very similar to theirs in that both their implementation and ours abstract from parameter estimation error. Thus, our approach can be viewed as a formalization of their approach, albeit with a different null hypothesis. An additional advantage of our approach is that the confidence bands that we propose are joint, not pointwise.

The previous discussion suggests that we could also apply our approach to likelihood-ratio (LR) tests based on the inverse normal transformation of the PITs. It is well known that, when the forecast density is correctly specified, an inverse normal transformation of the PITs (ζ_{t+h}) has a standard normal distribution (Berkowitz, 2001).[16] As noted in the literature, the latter approach has typically abstracted from parameter estimation uncertainty. When focusing on the traditional null hypothesis, H_0^{CS}, ignoring parameter estimation error leads to size distortions. Note that the size distortion is not only a small sample phenomenon, but persists asymptotically. The next result shows that, since parameter estimation error is maintained under our null hypothesis H_0, eq. (3), there is no need to correct the asymptotic distribution and the implied critical values of the likelihood ratio tests to account for parameter estimation error.

Corollary 3 (Inverse Normal Tests) *Let $\Phi^{-1}(.)$ denote the inverse of the standard normal distribution function. Under Assumptions 1,2 and H_0 in eq. (3): $\zeta_{t+1} \equiv \Phi^{-1}(z_{t+1})$ is iid$N(0,1)$.*

Thus, one could test for the correct specification of the density forecast by testing the absence of serial correlation and the correct specification of the moments of ζ_{t+h}. For example, the researcher could estimate an AR(1) model for ζ_{t+1} and test that the mean and the slope are both zero, and that the variance is one. The advantage of this approach is that it is informative regarding the possible causes underlying the misspecification of the density forecast and it may perform better in small samples. The disadvantage of the approach is

[16] González-Rivera and Yoldas (2012) provide an extension of this test to multivariate out-of-sample predictive densities.

that, unlike the κ_P and C_P tests, it focuses on specific moments of the distribution rather than the whole (non-parametric) cumulative distribution function.

Finally, note that our approach provides not only a rationale to the common practice of evaluating the correct specification of density forecasts using PITs without adjusting for parameter estimation error (Diebold et al., 1998), but also a methodology for implementing tests robust to the presence of serial correlation as well as dynamically misspecified models. This is a more general case and we consider it in the next section.

3.2 Multi-step-ahead Forecasts and Dynamic Mis-specification

When considering h-step-ahead forecasts, $h > 1$ and finite, as well as when $h = 1$ for models that are dynamically misspecified, an additional problem arises, as both of these cases involve serial correlation in the PITs.[17] Thus, we need to extend our results and allow the forecasts to be both serially correlated and potentially misspecified under the null hypothesis; that is, Assumption 2 does not hold.

Under dynamically misspecified null or when evaluating h-step-ahead conditional predictive densities, we show that $\Psi_P(r)$ weakly converges (considered as variables in the space $[0,1] \times \mathbb{R}$) to the Gaussian process $\Psi(.,.)$, with mean zero and an auto-covariance function that depends on the serial correlation in the PITs.

Theorem 4 (Correct Specification Tests under Serial Correlation) *Under Assumption 1 and H_0 in eq. (3): (i) $\{z_{t+h}\}_{t=R}^{T}$ is $U(0,1)$; (ii) $\Psi_P(r)$ weakly converges as a variable in the space $[0,1] \times \mathbb{R}$) to the Gaussian process $\Psi(.)$, with mean zero and auto-covariance function $E[\Psi(r_1)\Psi(r_2)] = \sigma(r_1, r_2)$, where $\sigma(r_1, r_2) = \sum_{d=-\infty}^{\infty} [F_d(r_1, r_2) - F(r_1) F(r_2)]$. Furthermore,*

$$\kappa_P \Rightarrow \sup_{r \in [0,1]} \Psi(r)^2,$$

$$C_P \Rightarrow \int_0^1 \Psi(r)^2 \, dr.$$

For a given estimate of $\sigma(r_1, r_2)$, the critical values of κ_P and C_P can be obtained via Monte Carlo simulations.[18] Note that, although in the case of dynamically misspecified models our method has less computation advantages relative to Corradi and Swanson (2006c), as the limiting distribution of our test too depends on nuisance parameters, still our test is

[17] In fact, h-step-ahead forecasts are serially correlated of order at least $(h-1)$.

[18] In the Monte Carlo as well as in the empirical application we obtain the critical values of our tests using Newey and West's (1987) HAC estimator for the covariance of the PITs.

different because it considers a different null hypotheses. In addition, given that we maintain parameter estimation under the null, our test has wider applicability, since it does not require designing the appropriate bootstrap that correctly mimics the contribution of parameter estimation error to the asymptotic distribution and prove its validity, which could be challenging in many instances.

However, there are several other solutions proposed in the literature that one could use within our approach as well. A first approach is to discard data by reducing the effective sampling rate to ensure an uncorrelated sample (Persson, 1974 and Weiss, 1973). This can be implemented in practice when models are dynamically correctly specified by creating sub-samples of predictive distributions that are at least h periods apart. However, this procedure may not be possible in small samples, since the sub-samples may significantly reduce the size of the sample. In those cases, one may implement the procedure in several uncorrelated sub-samples of forecasts that are at least h periods apart and then use Bonferroni methods to obtain a joint test without discarding observations (see Diebold et al., 1998). However, it is well-known that Bonferroni methods are conservative; thus the latter procedure, while easy to implement, may suffer from low power.

4 How to Use Our Tests?

Suppose the researcher decides to use the tests described in Theorem 2. If the tests reject, the rejection could be due to the violation of either independence or uniformity. If the researcher is concerned that the data may not be independent, he/she could use our test for uniformity robust to violations of independence. As discussed in Section 3.2, the latter test is not nuisance parameter free, so the implementation is more challenging and requires simulating the critical values. Alternatively, one could test for serial correlation in a way that is robust to uniformity. Among the tests that could be implemented, one could consider the Ljung-Box Q or Box-Pierce Q-test statistics (Box and Pierce, 1970) or the BDS test proposed by Brock, Dechert and Scheinkman (1987). The Q-test detects auto-correlation in a linear framework whereas the BDS test is a non-parametric test of independence and identical distribution against an unspecified alternative. Note that serial correlation implies lack of independence but serial uncorrelatedness does not necessarily imply independence. If the PITs are not serially correlated, the researcher should feel more comfortable in applying the critical values provided in our paper. Note, however, that in this case, independence could still be violated.

5 Monte Carlo Evidence

In this section we analyze the size and power properties of our proposed tests in small samples for both correctly specified and misspecified forecasting models. Note that comparisons with alternative methods (such as Corradi and Swanson, 2006c or González-Rivera and Yoldas, 2012), are not meaningful since we focus on a null hypothesis that is different from theirs.

5.1 Size Analysis

To investigate the size properties of our tests we consider several Data Generating Processes (DGPs). The forecasts are based on model parameters estimated in rolling windows for $t = R, ..., T + h$. We consider several values for in-sample estimation window of $R = [25, 50, 100, 200]$ and out-of-sample evaluation period $P = [25, 50, 100, 200, 500, 1000]$ to evaluate the performance of the proposed procedure. While our Assumptions require R finite, we consider both small and large values of R to investigate the robustness of our methodology when R is large. The DGPs are the following:

DGP S1 (Baseline Model): We estimate a model $y_t = \beta x_{t-1} + e_t$, $e_t \sim iidN(0,1)$. The data is generated by $y_t = \mu_t + x_{t-1} + \varepsilon_t$, $\varepsilon_t \sim iid\ N(0,1)$ and $x_t \sim iid\ N(0,1)$,where

$$\mu_t = \left(\frac{R^{-1}\sum_{j=t-R+1}^{t}\left[x_{j-1} - \left(R^{-1}\sum_{s=t-R+1}^{t} x_{s-1}\right)\right]\left[y_j - \left(R^{-1}\sum_{s=t-R+1}^{t} y_s\right)\right]}{R^{-1}\sum_{j=t-R+1}^{t}\left[x_{j-1} - \left(R^{-1}\sum_{s=t-R+1}^{t} x_{s-1}\right)\right]^2} - 1 \right) x_t.$$

DGP S2 (Extended Model): We parameterize the model according to the realistic situation where the researcher is interested in forecasting one-quarter-ahead U.S. real GDP growth with the lagged term spread from 1959:I-2010:III. We estimate a model $y_t = \beta x_{t-1} + e_t$, $e_t \sim iidN(0,1)$, while the data has been generated with the DGP: $y_t = \mu_t + \gamma x_{t-1} + \varepsilon_t$, $\varepsilon_t \sim iidN\left(0,1\right)$, $x_t = 0.2 + 0.8x_{t-1} + \nu_t$, $\nu_t \sim iid\ N(0, 1.08^2)$ independent from ε_t, $\gamma = 0.48$ and

$$\mu_t = \left(\frac{R^{-1}\sum_{j=t-R+1}^{t}\left[x_{j-1} - \left(R^{-1}\sum_{s=t-R+1}^{t} x_{s-1}\right)\right]\left[y_j - \left(R^{-1}\sum_{s=t-R+1}^{t} y_s\right)\right]}{R^{-1}\sum_{j=t-R+1}^{t}\left[x_{j-1} - \left(R^{-1}\sum_{s=t-R+1}^{t} x_{s-1}\right)\right]^2} - \gamma \right) x_t.$$

DGPs S1-S2 are based on one-step-ahead forecast densities. DGP S3 considers the case of h-step-ahead forecast densities where the PITs are serial correlated by construction.

DGP S3 (Serial Correlation): The DGP is $y_t = \mu_t + x_{t-1} + \varepsilon_t + \rho\varepsilon_{t-1,}$ $\varepsilon_t \sim iidN(0,1)$, $x_t \sim iid\ N(0,1)$, $\rho = 0.2$ and μ_t is as defined in DGP S1. The estimated model is: $y_t = \beta x_{t-1} + e_t, e_t \sim iid\ N(0, 1+\rho^2)$.

The results are shown in Table 2. The table shows that our tests performs well in finite samples, with mild under-rejections in DGP S2. In the case of serial correlation, DGP S3, the asymptotic distribution of the tests in Theorem 3 approximated using HAC-consistent variance estimates tends to over-reject, although mildly.[19]

INSERT TABLE 2 HERE

5.2 Power Analysis

To investigate the power properties of our tests, we consider the case of constant misspecification in the following DGP.

DGP P: The data are generated from a linear combination of normal and χ^2_1 distributions: $y_t = \mu_t + x_{t-1} + (1-c)\,\widehat{\sigma}_t \eta_{1,t} + c\left(\eta^2_{2,t} - 1\right)\sqrt{2}$, where $x_t, \eta_{1,t}$ and $\eta_{2,t}$ are $iidN\,(0,1)$ random variables that are independent of each other and μ_t is as defined in DGP S1. The researcher tests whether the data result from a normal distribution, i.e. considers the model $y_t = \beta x_{t-1} + e_t,\ e_t \sim iidN(0,\sigma_e)$. When c is zero, the null hypothesis is satisfied. When c is positive, the considered density becomes a convolution of a standard normal and a χ^2_1 distribution (with mean zero and variance one), where the weight on the latter becomes larger as c increases.[20]

The results shown in Table 3 suggest that our proposed specification tests (κ_P, C_P) have good power properties in detecting misspecification in the predictive density.[21]

INSERT TABLE 3 HERE

6 Empirical Analysis

This section provides an empirical assessment of the correct specification of widely-used density forecasts: the Survey of Professional Forecasters' (SPF) density forecasts of inflation and output growth, and density forecasts of the seven macroeconomic aggregates in a representative macroeconomic model.

[19] The size of the test might be improved by finite sample corrections. For instance, one could use a version of the block bootstrap suggested by Inoue (2001).

[20] Note that $\left(\eta^2_{2,t} - 1\right)\sqrt{2}$ is a chi-squared distribution with zero mean and variance one, that is, it has the same mean and variance as the normal distribution we have under the null hypothesis, although the shape is different.

[21] Unreported results show that the test still has power when we consider smaller sample sizes, e.g. $T = 100$.

6.1 Evaluation of SPF Density Forecasts

Diebold et al. (1999) evaluate the correct specification of the density forecasts of inflation in the SPF.[22] In this section, we conduct a formal test of correct specification for the SPF density forecasts using our proposed procedure and compare our results to theirs. In addition to inflation, we also investigate the conditional density forecasts of output growth.

We use real GNP/GDP and the GNP/GDP deflator as measures of output and prices. The mean probability distribution forecasts are obtained from the Federal Reserve Bank of Philadelphia (Croushore and Stark, 2001). In the SPF data set, forecasters are asked to assign a probability value (over pre-defined intervals) of year-over-year inflation and output growth for the current (nowcast) and following (one-year-ahead) calendar years. The forecasters update the assigned probabilities for the nowcasts and the one-year-ahead forecasts on a quarterly basis. The probability distribution provided by the SPF is discrete, and we base our results on a continuous approximation by fitting a normal distribution. The realized values of inflation and output growth are based on the real-time data set for macroeconomists, also available from the Federal Reserve Bank of Philadelphia.[23]

The analysis of the SPF probability distribution is complicated since the SPF questionnaire has changed over time in various dimensions: there have been changes in the definition of the variables, the intervals over which probabilities have been assigned, as well as the time horizon for which forecasts have been made. To mitigate the impact of these problematic issues, we truncate the data set and consider only the period 1981:III-2011:IV. We use the year-over-year growth rates of output and prices calculated from the first quarterly vintage of real GNP/GDP and the GNP/GDP deflator in each year to evaluate the density forecasts. For instance, in order to obtain the growth rate of real output for 1981, we take the 1982:I vintage of data and calculate the growth rate of the annual average GNP/GDP from 1980 to 1981. We consider the annual-average over annual-average percent change (as opposed to fourth-quarter over forth-quarter percent change) in output and prices to be consistent with the definition of the variables that SPF forecasters provide probabilistic predictions for.

The empirical results are shown in Table 4. Asterisks ('*') indicate rejection at the 5% significance level based on the critical values in Theorem 2 (reported in Table 1, Panel A),

[22] The SPF provides two types of density forecasts: one is the distribution of point forecasts across forecasters (which measures the dispersion of point forecasts across forecasters), and the other is the mean of the probability density forecasts (which measures the average of the density forecasts across forecasters). We focus on the latter.

[23] The data are available at http://www.philadelphiafed.org/research-and-data/real-time-center.

while '†' indicates rejection at 5% significance level based on the critical values in Theorem 4. The latter are simulated conditional on the data (i.e. conditional on the HAC estimate of the variance-covariance matrix of the PIT, with a truncation parameter equal to 3). The tests which are robust to violations of independence (based on Theorem 4) as well as the ones that maintain independence under the null reject correct specification for both output growth and inflation, except for output growth at the one-year-ahead forecast horizon. The test robust to the violation of independence under the null favors correct specification for the current year forecast of the output growth as well.

INSERT TABLE 4 HERE

Our results are important in light of the finding that survey forecasts are reportedly providing the best forecasts of inflation. For example, Ang et al. (2007) find that survey forecasts outperform other forecasting methods (including the Phillips curve, the term structure and ARIMA models) and that, when combining forecasts, the data put the highest weight on survey information. Our results imply that survey forecasts still do not provide correct forecasts for the whole distribution of inflation.

Figure 2 plots the empirical CDF of the PITs (solid line). Under the null hypothesis in Theorem 2, the PITs should be uniformly distributed; thus the CDF of the PITs should be the 45 degree line. The figure also reports the critical values based on the κ_P test. If the empirical CDF of the PITs is outside the critical value lines, we conclude that the density forecast is misspecified. Clearly, the correct specification is rejected in all cases except the one-year-ahead density forecast of GDP growth. The figure also provides a visual analysis of the misspecification in the PITs: the survey typically overpredicts future large realizations (both positive and negative) of output growth and inflation.

For comparison, Figure 3 reports results based on Diebold et al.'s (1998) test. Panel A plots the empirical distribution of the PITs of output growth for both the density nowcast (left-hand panel) and the one-year-ahead density forecast (right-hand panel). In addition to the PITs, we also provide the 95% confidence interval (dotted lines) using a normal approximation to a binomial distribution similar to Diebold et al.'s (1998). Both nowcast and one-year-ahead density forecasts of output growth are misspecified, although misspecification is milder in the case of one-year-ahead output growth. Figure 3, Panel B, shows the PITs for inflation. According to this test, both the density nowcast and one-year-ahead forecast overestimate tail risk. This phenomenon is more pronounced for the nowcast. Overall, the results obtained by using Diebold et al.'s (1998) test are broadly similar to those obtained

by using the test that we propose in this paper, with one important exception. In the case of one-year-ahead GDP growth forecasts, our test based on Theorem 2 does not reject, whereas the Diebold et al. (1998) test does, despite the fact that both rely on *iid* assumptions. The discrepancy in the results is most likely due to the fact that the latter test is pointwise, whereas we jointly test the correct specification across all quantiles in the empirical distribution function: thus our test has larger critical values than the latter, in order to correctly account for the joint null hypothesis.

INSERT FIGURES 2 AND 3 HERE

6.2 Evaluation of a Baseline Macroeconomic Model

Macroeconomic models are widely used in central banks for policy evaluation and forecasting. Several recent contributions have focused on the ability of Dynamic Stochastic General Equilibrium (DSGE) models to produce good out-of-sample point forecasts. In particular, Smets and Wouters (2007) show that the forecasts of their model that they propose are competitive relative to Bayesian VAR forecasts. Edge, Kiley and Laforte (2010) evaluate the predictive ability of the Federal Reserve Board's model (Edo), and Edge and Gürkaynak (2010) provide a thorough analysis of the forecasting ability of the same model using real-time data. The main result in the latter is that point forecasts of macroeconomic models perform similarly to that of a constant mean model, but both are biased; the reason why they perform similarly is because volatility was low during the Great Moderation sample period they consider, and, therefore, most variables were unpredictable. Edge, Gürkaynak and Kısacıkoğlu (2013) extend the results of Edge and Gürkaynak (2010) to a longer sample and Gürkaynak, Kısacıkoğlu and Rossi (2013) analyze the point forecasting ability of the models relative to reduced-form models, and find that the latter perform better than the DSGE model at some forecast horizons.

While the contributions discussed above focus on evaluating how accurate macroeconomic models' point forecasts are, central banks are becoming more and more interested in analyzing the uncertainty around the point forecasts that macroeconomic models provide. In this section, we focus on evaluating density forecasts of a baseline DSGE model, a task that only a few recent contributions have performed. Christoffel, Coenen and Warne (2010) study the performance of density forecasts of the European Central Bank's model (NAWM) and find that it tends to overestimate nominal wages. Wolters (2012) evaluates point and density forecasts for US inflation and concludes that the models overestimate uncertainty

around point forecasts. Bache, Jore, Mitchell and Vahey (2011) combine density forecasts of inflation from VARs and a macroeconomic model using the linear opinion pool. They find that allowing for structural breaks in the VAR produces well-calibrated density forecasts for inflation but reduces the weight on the macroeconomic model considerably. Our paper differs from the literature as we evaluate the model-based density forecasts using our novel PIT-based test and compare its results with those based on the PIT-based tests proposed by Diebold et al. (1998).

We focus on the Smets and Wouters (2007) model as our benchmark model. The model is a real business cycle model with both nominal as well as real rigidities; in fact, it features sticky prices and wages as well as habit formation in consumption and cost of adjustment in investment.[24] We recursively re-estimate the model (using exactly the same data and priors) in fixed rolling window of 80 observations and produce a sequence of 80 out-of-sample density forecasts.[25] The model includes seven observables and seven shocks; we separately evaluate the forecast densities for each of the target variables. We focus on the one-quarter-ahead forecast horizon.[26]

Table 5 reports the empirical results for the correct calibration of the model's density forecasts. The last two columns report the value of the κ_P and C_P tests that we propose in this paper. Asterisks '*' indicate rejection at the 5% significance level based on the critical values in Theorem 2 (reported in Table 1, Panel A), while '†' indicates rejection at 5% significance level based on the critical values in Theorem 4 with a HAC estimate of the covariance matrix. According to the critical values in Theorem 2, the density forecasts of investment, inflation, hours and wages are well calibrated, although those of the remaining variables are not. When one allows for serial correlation under the null (that is, using the critical values implied by Theorem 4), then invesment, hours and wages pass the test of correct calibration. Since the Ljung-Box test rejects that the PITs are uncorrelated for all

[24] See Section I in Smets and Wouters (2007) for a detailed description of the model.

[25] Smets and Wouters (2007) approximate the deciles of the predictive densities based on Gaussian kernel estimates, given the DSGE's assumption of normally distributed errors. We obtain the PITs using a linear interpolation for the inter-decile range.

[26] The sample period is from 1966:I to 2004:IV. The first one-quarter-ahead out-of-sample forecast is for 1985:I. From the 80 observations in each rolling window, 4 are used for pre-sampling: they are not included in the likelihood. The total number of out-of-sample periods is 80. The model is estimated using Dynare. We create a sample of 150,000 draws for each rolling window estimation, discarding the first 20% of the draws. We use a step-size of 0.2 for the jumping distribution in the Metropolis-Hastings algorithm, resulting in rejection rates hovering around 0.4 across various estimation windows.

variables besides consumption and inflation (at least in one of the moments – see the results reported in the same table) at 5% significance level and all of them at 10% significance level, the version of our test that maintains serial correlation under the null is the appropriate one.

Figure 4 displays the cumulative distribution of the PITs for each the observables, together with critical values for correct calibration based on the κ_P test in Theorem 2. The figures show that there are too few realizations of consumption, output growth and the federal funds rate in the lowest quantiles of the distribution; that is, the model overpredicts the lower tail values of the target variable. For the remaining variables, the test suggests proper calibration. For comparison, Figure 5 shows the estimated PDF of the PITs, together with critical values based on Diebold et al. (1998). Diebold et al.'s (1998) methodology produces results similar to ours, except for hours worked and real wage forecasts, which are correctly specified according to our test and misspecified according to Diebold et al.'s (1998) test.[27] Again, the most likely reason for the discrepancy appears to be the different nature of the test: our test is joint across deciles whereas the latter is pointwise.

INSERT FIGURES 4 AND 5 HERE

7 Conclusions

This paper proposes new tests for predictive density evaluation. The techniques are based on Kolmogorov-Smirnov and Cramér-von Mises-type test statistics. We provide critical values of the tests for dynamically correctly specified models as well as tests that focus on specific parts of the predictive density. We also propose methodologies that can be applied to dynamically misspecified models and multiple-step-ahead forecast horizons. Our empirical analyses uncover that both SPF output growth and inflation density forecasts as well as DSGE-based forecasts of several macroeconomic aggregates are misspecified.

[27] Note that this is a fair comparison, since both Figures 4 and 5 are constructed under the maintained assumption of independence.

References

[1] Aastveit, K.A., C. Foroni and F. Ravazzolo (2014), "Density forecasts with MIDAS models", *mimeo.*

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

[3] Ang, A., G. Bekaert and M. Wei (2007), "Do Macro Variables, Asset Markets or Surveys Forecast Inflation Better?" *Journal of Monetary Economics* 54, 1163-1212.

[4] Bache, I.W., A.S. Jore, J. Mitchell and S.P. Vahey (2011), "Combining VAR and DSGE Forecast Densities," *Journal of Economic Dynamics and Control* 35, 1659-1670.

[5] Bai, J. (2003), "Testing Parametric Conditional Distributions of Dynamic Models," *Review of Economics and Statistics* 85(3), 531-549.

[6] Bai, J. and S. Ng (2005), "Tests for Skewness, Kurtosis, and Normality for Time Series Data," *Journal of Business and Economic Statistics* 23(10), 49-60.

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

[8] Billio, M., R. Casarin, F. Ravazzolo and H. van Dijk (2013), "Time-varying Combinations of Predictive Densities using Nonlinear Filtering", *Journal of Econometrics* 177(2), 213–232.

[9] Bontemps, C. and N. Meddahi (2012), "Testing Distributional Assumptions: A GMM Approach," *Journal of Applied Econometrics* 27(6), 978-1012.

[10] Box, G. and D. Pierce (1970), "Distribution of Residual Auto-correlation in Autoregressive-Integrated Moving Average Time Series Models," *Journal of the American Statistical Association* 65, 1509-1526.

[11] Brock, W. A., W. Dechert and J. Scheinkman (1987), "A Test for Independence based on the Correlation Dimension," *Working Paper,* University of Winconsin at Madison, University of Houston, and University of Chicago.

[12] Christoffel, K., G. Coenen and A. Warne (2010), "Forecasting with DSGE Models," *ECB Working paper* 1185.

[13] Clements, M. P. (2004), "Evaluating the Bank of England Density Forecasts of Inflation," *The Economic Journal* 114, 844–866.

[14] Corradi, V. and N. R. Swanson (2006a), "Bootstrap Conditional Distribution Tests in the Presence of Dynamic Misspecification," *Journal of Econometrics* 133, 779-806.

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

[16] Corradi, V. and N. R. Swanson (2006c), "Predictive Density and Conditional Confidence Interval Accuracy Tests," *Journal of Econometrics* 135(1–2), 187-228.

[17] Croushore, D. and T. Stark (2001), "A Real-time Data Set for Macroeconomists," *Journal of Econometrics* 105(1), 111-130.

[18] Davidson, J. (1994), *Stochastic Limit Theory: An Introduction for Econometricians*, Oxford University Press.

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

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

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

[22] Edge, R. M. and R. S. Gürkaynak (2010), "How Useful Are Estimated DSGE Model Forecasts for Central Bankers?" *Brookings Papers on Economic Activity* 41(2), 209-259.

[23] Edge, R. M., R. S. Gürkaynak, and B. Kısacıkoğlu (2013), "Judging the DSGE Model by Its Forecast," *mimeo*.

[24] Edge, R. M., M. T. Kiley and J. P. Laforte (2010), "A Comparison of Forecast Performance Between Federal Reserve Staff Forecasts, Simple Reduced-form Models, and a DSGE Model," *Journal of Applied Econometrics* 25(4), 720-754.

[25] Elliott, G. and A. Timmermann (2008), "Economic Forecasting", *Journal of Economic Literature* 46, 3-56.

[26] Elliott, G. and A. Timmermann, *Economic Forecasting*, Princeton University Press, forthcoming.

[27] Franses, P. H. and D. van Dijk (2003), "Selecting a Nonlinear Time Series Model using Weighted Tests of Equal Forecast Accuracy," *Oxford Bulletin of Economics and Statistics* 65, 727–744.

[28] Giacomini, R. and H. White (2006), "Tests of Conditional Predictive Ability", *Econometrica* 74(6), 1545-1578.

[29] González-Rivera, G. and E. Yoldas (2012), "Autocontour-based evaluation of Multivariate Predictive Densities," *International Journal of Forecasting* 28(2), 328-342.

[30] González-Rivera, G. and Y. Sun (2014), "Generalized Autocontours: Evaluation of Multivariate Density Models," *International Journal of Forecasting*, forthcoming.

[31] Gürkaynak, R. S., B. Kisacikoglu and B. Rossi (2013), "Do DSGE Models Forecast More Accurately Out-of-Sample than VAR Models?" In: T. Fomby, L. Kilian and A. Murphy (eds.), *Advances in Econometrics: VAR Models in Macroeconomics –New Developments and Applications Vol. 31*, forthcoming.

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

[33] Hong, Y. M. and H. Li (2005), "Nonparametric Specification Testing for Continuous Time Models with Applications to Term Structure of Interest Rates," *Review of Financial Studies* 18(1), 37-84.

[34] Hong, Y., H. Li and F. Zhao (2007), "Can the Random Walk Model Be Beaten in Out-of-sample Density Forecasts? Evidence From Intraday Foreign Exchange Rates," *Journal of Econometrics* 141(2), 736–776.

[35] Inoue, A. (2001), "Testing for Distributional Change in Time Series," *Econometric Theory* 17, 156-187.

[36] Inoue A. and B. Rossi (2012), "Out-of-sample Forecast Tests Robust to the Window Size Choice," *Journal of Business and Economics Statistics* 30(3), 432-453.

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

[38] Newey, W.K. and K.D. West (1987), "A Simple, Positive semi-definite, Heteroskedasticity and Auto-correlation Consistent Covariance Matrix," *Econometrica* 55(3), 703-708.

[39] Persson, J. (1974), "Comments on Estimations and Tests of EEG Amplitude Distributions," *Electroencephalography and Clinical Neurophysiology* 37, 309-313.

[40] Ravazzolo, F. and S. P. Vahey (2014), "Forecast Densities for Economic Aggregates from Disaggregate Ensembles", *Studies of Nonlinear Dynamics and Econometrics* 18(4), 367–381.

[41] Shorack, G. R. and J. A. Wellner (1986), *Empirical Processes with Applications to Statistics*, Wiley.

[42] Smets, F. and R. Wouters (2007), "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review* 97(3), 586-607.

[43] White, H. (2001), *Asymptotic Theory for Econometricians,* Revised Edition, Academic Press.

[44] Wolters, M. H. (2012), "Evaluating Point and Density Forecasts of DSGE Models," *MPRA Working Paper* 36147.

[45] Weiss, M. S. (1973), "Modifications of the Kolmogorov-Smirnov Statistic for Use with Correlated Data," *Journal of the American Statistical Association* 74, 872-875.

Appendix A. Proofs

The appendix provides the proofs for Theorems 1, 2, 4 and Corollary 4.

Proof of Theorem 1. (i) The true joint conditional predictive density of $\{y_{t+1}\}_{t=R}^{T}$ can be decomposed as $\phi_0(y_{T+1},...,y_R|\mathcal{F}_R) = \phi_0(y_{T+1}|\mathcal{F}_T)\,\phi_0(y_T|\mathcal{F}_{T-1})...\phi_0(y_{R+1}|\mathcal{F}_R)$, where R is finite by Assumption 1(iv) (which guarantees that we condition on a finite information set). Let $q(z_{R+1},...,z_{T+1}|\mathcal{F}_R)$ denote the conditional joint density of the probability integral transforms. Then, given that $y_{t+1} = \Phi_{t+1}^{-1}(z_{t+1}|\Im_t)$, where $\Im_t \subseteq \mathcal{F}_t$, we can re-write the joint density of the PITs as $q(z_{R+1},...,z_{T+1}|\mathcal{F}_R) = \phi_0\left(\Phi_{R+1}^{-1}(z_{R+1}|\Im_R)|\mathcal{F}_R\right)...\phi_0\left(\Phi_T^{-1}(z_T|\Im_{T-1})|\mathcal{F}_{T-1}\right) \times ... \times \phi_0\left(\Phi_{T+1}^{-1}(z_{T+1}|\Im_T)|\mathcal{F}_T\right)$.

By using the change of variables formula and Assumption 2(b),

$$
\begin{aligned}
q(z_{R+1},...,z_{T+1}|\mathcal{F}_R) = & \begin{vmatrix} \left(\partial\Phi_{R+1}^{-1}(z_{R+1}|\Im_R)/\partial z_{R+1}\right) & ... & \left(\partial\Phi_{R+1}^{-1}(z_{R+1}|\Im_R)/\partial z_{T+1}\right) \\ ... & ... & ... \\ \left(\partial\Phi_{T+1}^{-1}(z_{T+1}|\Im_T)/\partial z_{R+1}\right) & ... & \left(\partial\Phi_{T+1}^{-1}(z_{T+1}|\Im_T)/\partial z_{T+1}\right) \end{vmatrix} \\
& \times \phi_0\left(\Phi_{R+1}^{-1}(z_{R+1}|\Im_R)|\mathcal{F}_R\right)...\phi_0\left(\Phi_T^{-1}(z_T|\Im_{T-1})|\mathcal{F}_{T-1}\right)\phi_0\left(\Phi_{T+1}^{-1}(z_{T+1}|\Im_T)|\mathcal{F}_T\right) \\
& = \left(1/\phi_{R+1}(y_{R+1}|\Im_R)\right)...\left(1/\phi_T(y_T|\Im_{T-1})\right)\left(1/\phi_{T+1}(y_{T+1}|\Im_T)\right) \times \\
& \times \phi_0(y_{R+1}|\mathcal{F}_R)...\phi_0(y_T|\mathcal{F}_{T-1})\,\phi_0(y_{T+1}|\mathcal{F}_T),
\end{aligned}
$$

where the last equality holds because the Jacobian is lower triangular provided we are in a conditional forecasting framework and thus $\{y_{t+1,},...,y_{T+1}\} \notin \Im_t$ at any time t. Then,

$$
q(z_{R+1},...,z_{T+1}|\mathcal{F}_R) = \frac{\phi_0(y_{R+1}|\mathcal{F}_R)}{\phi_{R+1}(y_{R+1}|\Im_R)} \times ... \times \frac{\phi_0(y_T|\mathcal{F}_{T-1})}{\phi_T(y_T|\Im_{T-1})} \times \frac{\phi_0(y_{T+1}|\mathcal{F}_T)}{\phi_{T+1}(y_{T+1}|\Im_T)}.
$$

Now suppose that $\Im_t = \Im_{t-R+1}^t$, where $\Im_{t-R+1}^t$ contains only data available from time $t-R+1$ to time t. In other words, $\Im_{t-R+1}^t$ differs from $\Im_t$ because the rolling window does not use all the available information in the sample. If $\phi_{t+1}(y_{t+1}|\Im_t) = \phi_{t+1}\left(y_{t+1}|\Im_{t-R+1}^t\right) = \phi_{t+1}(y_{t+1}|\mathcal{F}_t)$ (the condition imposed by Assumption 2(a)), i.e. when $\Im_{t-R+1}^t$ contains all relevant past information, then we could re-write the above as

$$
q(z_{R+1},...,z_{T+1}|\mathcal{F}_R) = \frac{\phi_0(y_{R+1}|\mathcal{F}_R)}{\phi_{R+1}(y_{R+1}|\Im_1^R)} \times ... \times \frac{\phi_0(y_T|\mathcal{F}_{T-1})}{\phi_T\left(y_T|\Im_{T-R}^{T-1}\right)} \times \frac{\phi_0(y_{T+1}|\mathcal{F}_T)}{\phi_{T+1}\left(y_{T+1}|\Im_{T-R+1}^T\right)}
$$

It follows that, under the null, each ratio yields a $U(0,1)$ variable (since the PDF is the unit line), thus the joint distribution is a multivariate $U(0,1)$. In addition, since the joint distribution is the product of the marginals, then $\{z_{t+1}\}_{t=R}^{T}$ is $iid\ U(0,1)$.

(ii) Under H_0, z_{t+1} is uniformly distributed on $[0,1]$. The result follows from Theorem 4 noting that, from Inoue (2001 p.161, letting $r=1$ in his notation), under iid, the covariance simplifies to $\sigma(r_1,r_2) = F_0(r_1,r_2) - F(r_1)F(r_2) = \min(r_1,r_2) - r_1 r_2$, where the last equality follows from Shorack and Wellner (1986, p.131) and the fact that $\{z_{t+1}\}_{t=R}^{T}$ is uniform. ■

Proof of Theorem 2. The theorem follows from Theorem 1 by the Continuous Mapping theorem. ■

Proof of Corollary 3. The theorem follows directly from part (i) in Theorem 1 and Berkowitz (2001). ■

Proof of Theorem 4. (i) Under Assumption 1(ii) and H_0 in eq. (3), by Lemma 1 in Bai (2003), $\{z_{t+h}\}_{t=R}^{T}$ is $U(0,1)$. (ii) Clearly, Assumption 1(i) satisfies assumption (i) in Theorem 1 in Giacomini and White (2005), as they require strong mixing of the same size, with $\lambda > 1$. If Z_t is strong mixing with coefficients of size $\alpha(m)$, so is any measurable function of Z_t such that $g(Z_t, ..., Z_{t-R})$, where R is finite (White, 2001, Theorem 3.49); in our context, $g(Z_t, ..., Z_{t-R})$ is the cumulative distribution function and R is finite by Assumption 1(iv). Furthermore, in what follows, we show that (i) also satisfies Inoue's (2001) Assumption A. Since $g(Z_t, ..., Z_{t-R})$ is strong mixing with $\alpha(m)$ of size $-\frac{\lambda}{\lambda-1}$ then $\alpha(m) = O(m^{-\frac{\lambda}{\lambda-1}-\epsilon})$ for some $\epsilon > 0$ (White, 2001, Definition 3.45). That is, there exists a constant $B < \infty$ such that $\frac{|\alpha(m)|}{m^{-\frac{\lambda}{\lambda-1}-\varepsilon}} \leq B$ for every m (Davidson, 1994, p.31). Assumption A in Inoue (2001) requires that $\sum_{m=1}^{\infty} m^2 \alpha(m)^{\frac{\gamma}{4+\gamma}} < \infty$ for some $\gamma \in (0,2)$. Note that

$$\begin{aligned}\sum_{m=1}^{\infty} m^2 \alpha(m)^{\frac{\gamma}{4+\gamma}} &\leq \sum_{m=1}^{\infty} m^2 \left|\frac{\alpha(m)}{m^{-\frac{\lambda}{\lambda-1}-\varepsilon}}\right|^{\frac{\gamma}{4+\gamma}} m^{-\frac{\lambda}{\lambda-1}-\varepsilon} \\ &\leq B^{\frac{\gamma}{4+\gamma}} \sum_{m=1}^{\infty} m^2 m^{-\frac{\lambda}{\lambda-1}-\varepsilon} \leq \overline{B} \sum_{m=1}^{\infty} m^{2-\frac{\lambda}{\lambda-1}},\end{aligned}$$

where $\overline{B} \equiv B^{\frac{\gamma}{4+\gamma}} < \infty$. The series $\sum_{m=1}^{\infty} m^{2-\frac{\lambda}{\lambda-1}}$ is a harmonic series, convergent if $2 - \frac{\lambda}{\lambda-1} < -1$, i.e. if $\lambda < 3/2$. Thus, our Assumption 1(i) satisfies Inoue's Assumption A. Assumption 1(ii, iii) satisfy Inoue's Assumption B under the null. Consequently, Theorem 4 follows from Inoue (2001) by letting (in Inoue's notation) $r = 1$. ■

Appendix B. Tables and Figures

Table 1. Critical Values

		$\kappa_{\alpha;P}$			$C_{\alpha;P}$		
	$\alpha:$	0.01	0.05	0.10	0.01	0.05	0.10
Panel A. Tests on the Whole Distribution							
Correct Specification Test		2.50	1.72	1.39	0.77	0.47	0.35
Panel B. Tests on Specific Parts of the Distribution							
Right Tail	$r \in [0, 0.25]$	1.54	0.95	0.72	0.60	0.35	0.25
Right Half	$r \in [0, 0.50]$	2.42	1.54	1.21	0.90	0.53	0.39
Left Half	$r \in [0.50, 1]$	2.28	1.53	1.20	0.85	0.53	0.39
Left Tail	$r \in [0.75, 1]$	1.52	0.97	0.74	0.58	0.35	0.26
Center	$r \in [0.25, 0.75]$	2.52	1.73	1.36	1.21	0.72	0.52
Tails	$r \in \{[0, 0.25] \cup [0.75, 1]\}$	1.65	1.19	0.93	0.42	0.28	0.22

Note: Panel A reports critical values for the test statistics κ_P and C_P at the 1%,5% and 10% nominal sizes ($\alpha = 0.01, 0.05$ and 0.10). Panel B reports critical values for the same statistics for specific parts of the distributions, indicated in the second column. The number of Monte Carlo replications is 5,000. The domain for r is discretized with a lower bound of 0.01, upper bound of 0.99 and a step size of 0.005.

Table 2: Size Properties

DGP S1 (IID Case)									
		κ_P				C_P			
P	$R:$	25	50	100	200	25	50	100	200
25		0.047	0.045	0.048	0.045	0.046	0.047	0.049	0.046
50		0.053	0.052	0.052	0.053	0.053	0.050	0.054	0.052
100		0.045	0.044	0.049	0.049	0.048	0.046	0.045	0.052
200		0.053	0.050	0.052	0.054	0.049	0.051	0.052	0.049
500		0.049	0.053	0.059	0.052	0.048	0.052	0.050	0.050
1000		0.053	0.048	0.046	0.056	0.051	0.048	0.048	0.055
DGP S2 (IID Case)									
		κ_P				C_P			
P	$R:$	25	50	100	200	25	50	100	200
25		0.066	0.049	0.048	0.046	0.067	0.052	0.049	0.046
50		0.062	0.049	0.042	0.047	0.062	0.047	0.047	0.053
100		0.059	0.033	0.036	0.044	0.058	0.031	0.039	0.043
200		0.055	0.036	0.033	0.043	0.055	0.030	0.026	0.039
500		0.061	0.030	0.032	0.030	0.054	0.026	0.025	0.023
1000		0.053	0.033	0.024	0.024	0.050	0.025	0.020	0.022
DGP S3 (Serially Correlated Case)									
		κ_P				C_P			
P	$R:$	25	50	100	200	25	50	100	200
25		0.126	0.127	0.121	0.125	0.149	0.142	0.123	0.129
50		0.110	0.118	0.101	0.088	0.088	0.126	0.081	0.083
100		0.087	0.090	0.101	0.089	0.104	0.092	0.095	0.104
200		0.087	0.074	0.098	0.083	0.085	0.086	0.099	0.092
500		0.085	0.085	0.075	0.082	0.090	0.101	0.079	0.091
1000		0.088	0.096	0.101	0.084	0.101	0.101	0.101	0.088

Note: The table reports empirical rejection frequencies for the test statistics κ_P and C_P in eqs. (4) and (5) at the 5% nominal size for various values of P and R. The number of Monte Carlo replications is 5,000. The domain for r is discretized with a lower bound of 0.01, upper bound of 0.99 and a step size of 0.005. Critical values for DGP S1 and DGP S2 are those reported in Table 1, Panel A. For DGP S3, the critical values are simulated with a HAC estimate of a covariance matrix of the PITs.

Table 3. Power Properties

	DGP P	
c	κ_P	C_P
0	0.059	0.060
0.15	0.064	0.062
0.30	0.076	0.075
0.35	0.135	0.136
0.40	0.260	0.253
0.45	0.526	0.511
0.50	0.855	0.861
0.60	1.000	1.000

Note: The table reports empirical rejection frequencies for the test statistics κ_P and C_P in eqs. (4) and (5) for P=960 and R=40; the nominal size is 5%. The number of Monte Carlo replications is 5,000. The domain for r is discretized with a lower bound of 0.01, upper bound of 0.99 and a step size of 0.005. Critical values are those reported in Table 1, Panel A.

Table 4: Correct Specification Tests for SPF's Probability Forecasts

Series Name:	GDP Growth		GDP Deflator Growth	
Forecast Horizon (in rows):	κ_P	C_P	κ_P	C_P
0	2.31*	0.79*†	15.83*†	5.15*†
1	0.65	0.11	24.90*†	10.44*†

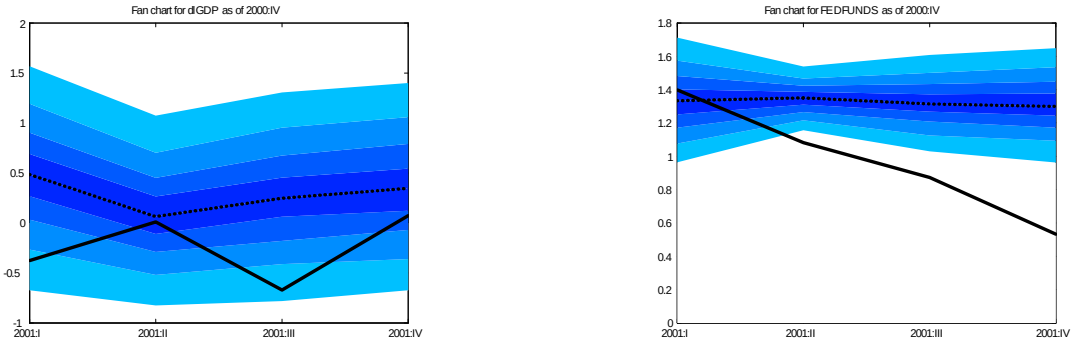
Note: Asterisks '*' indicate rejection at 5% significance level based on the critical values in Theorem 2 (reported in Table 1, Panel A), while '†' indicates rejection at 5% significance level based on the critical values in Theorem 4. The latter are simulated conditional on the data (i.e. conditional on the variance-covariance matrix of the PIT). The domain for r is discretized with a lower bound of 0.01, upper bound of 0.99 and a step size of 0.005.

Table 5: Correct Specification Tests for Model Forecast Distribution

Variable	LB: $(z_t - \bar{z})$	LB: $(z_t - \bar{z})^2$	κ_P	C_P
Consumption (real)	0.09	0.61	4.80 *†	1.94 *†
Investment (real)	0.00 *	0.52	0.48	0.15
Output Growth (real)	0.00 *	0.28	2.05 * †	0.67 *†
Inflation	0.07	0.83	1.51	0.46 †
Hours	0.00 *	0.53	0.88	0.31
Wages (real)	0.82	0.04 *	1.25	0.26
Federal Funds Rate	0.00 *	0.00 *	3.44 *†	1.41 *†

Note: The column labeled "LB" indicates p-values of the Ljung-Box test statistic for absence of serial correlation; values marked by '*' indicate rejections at 5% significance level. For the κ_P and C_P tests, '*' indicates rejection at the 5% significance level based on the critical values in Theorem 2 (reported in Table 1, Panel A), while '†' indicates rejection at 5% significance level based on the critical values in Theorem 4. The domain for r is discretized with a lower bound of 0.01, upper bound of 0.99 and a step size of 0.005. The evaluation sample is from 1985:I - 2004:IV.

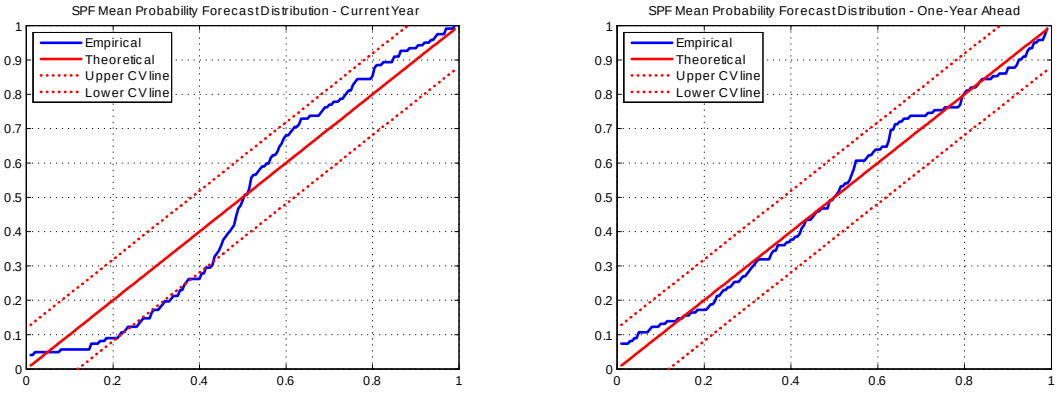
Figure 1. Representative Fan Charts from the Macroeconomic Model in 2000:IV



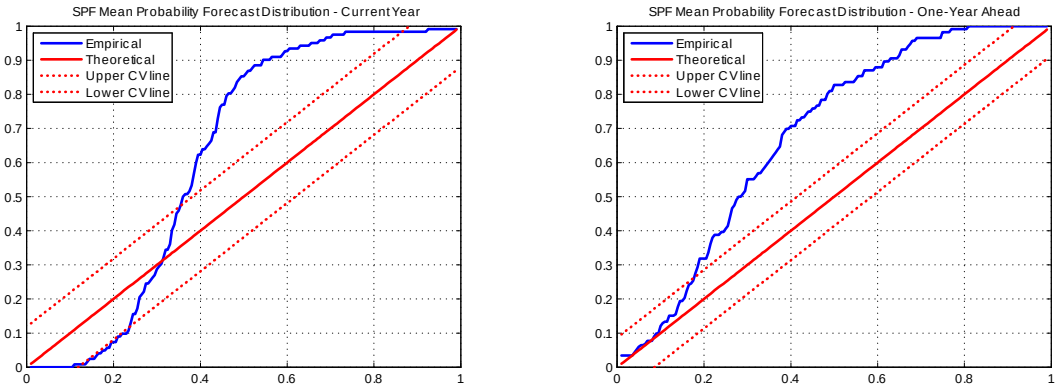
Note: The figure shows fan charts obtained by estimating the baseline model with data up to 2000:IV, prior to 2001:I-2001:IV recession. Depicted are the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th deciles of the predictive distribution for one to four-quarter-ahead out-of-sample forecasts. The solid lines represent the actual realizations of the data, while the dotted lines represent the median forecast.

Figure 2. CDF of the PITs – SPF Probability Forecast

Panel A: GDP Growth (1981:III-2011:IV)



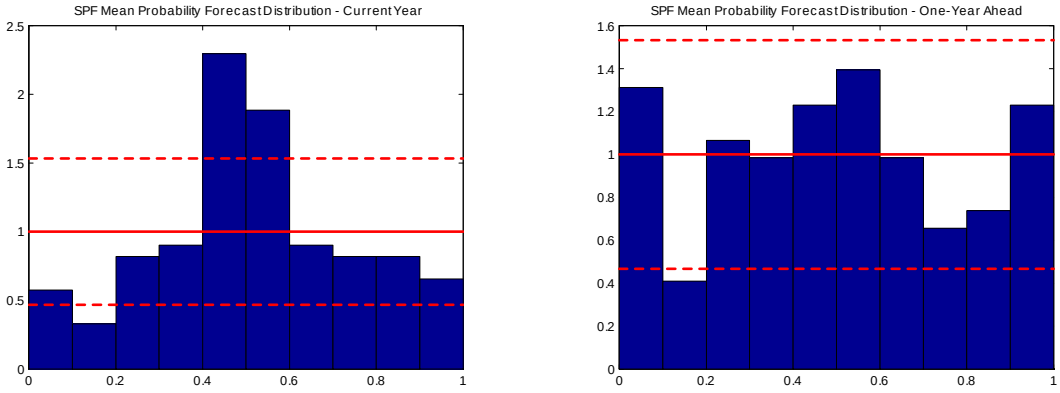
Panel B: GDP Deflator Growth (1981:III-2009:IV)



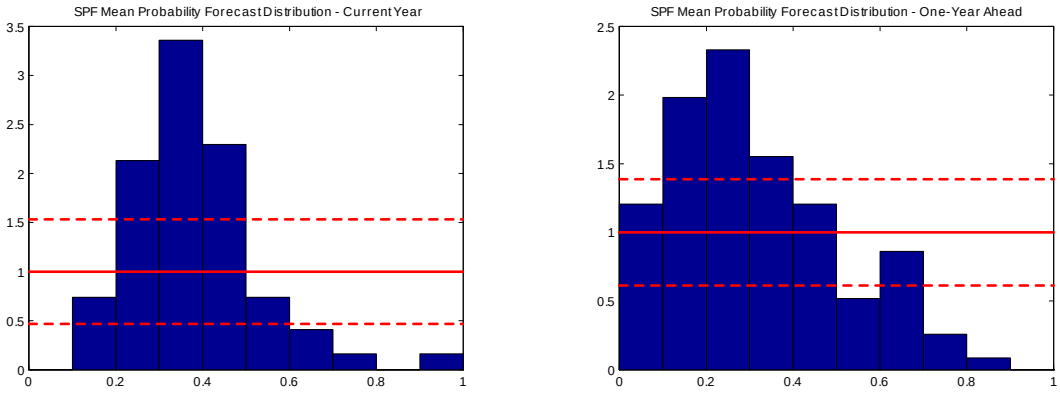
Note: The figure shows the empirical CDF of the PITs (solid line), the CDF of the PITs under the null hypothesis (the 45 degree line) and the 95% critical values based on the κ_P test reported in Table 1, Panel A.

Figure 3. PDF of the PITs – SPF Probability Forecast

Panel A: GDP Growth (1981:III-2011:IV)

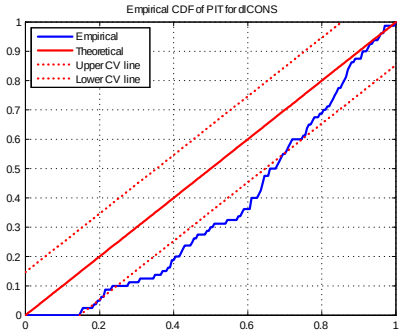


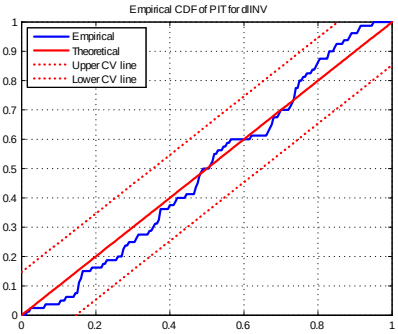
Panel B: GDP Deflator Growth (1981:III-2009:IV)

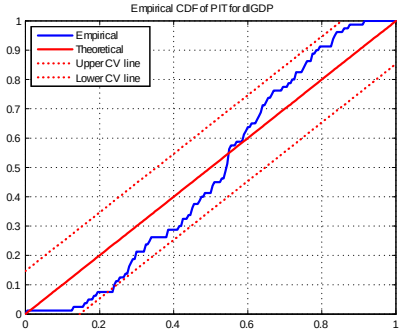


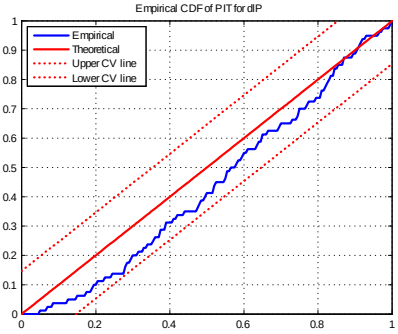
Note: The figures show the PDF of the PITs (normalized) and the 95% critical values approximated under Diebold et al.'s (1998) binomial distribution (dashed lines), constructed using a normal approximation.

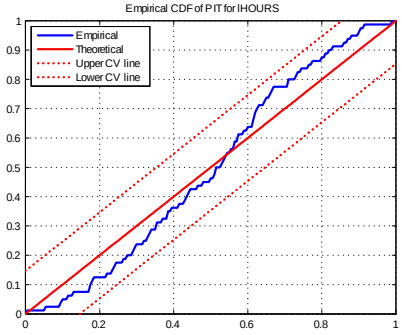
Figure 4. CDF of the PITs – Model Forecast Distribution (1985:I-2004:IV)

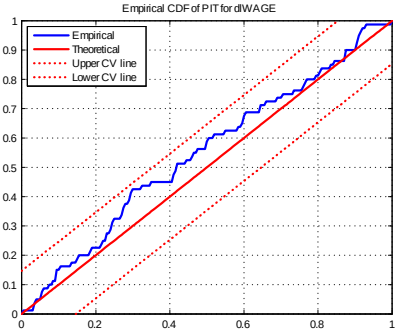


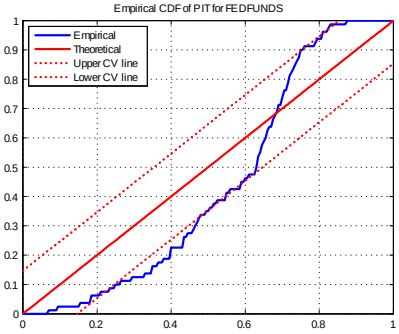






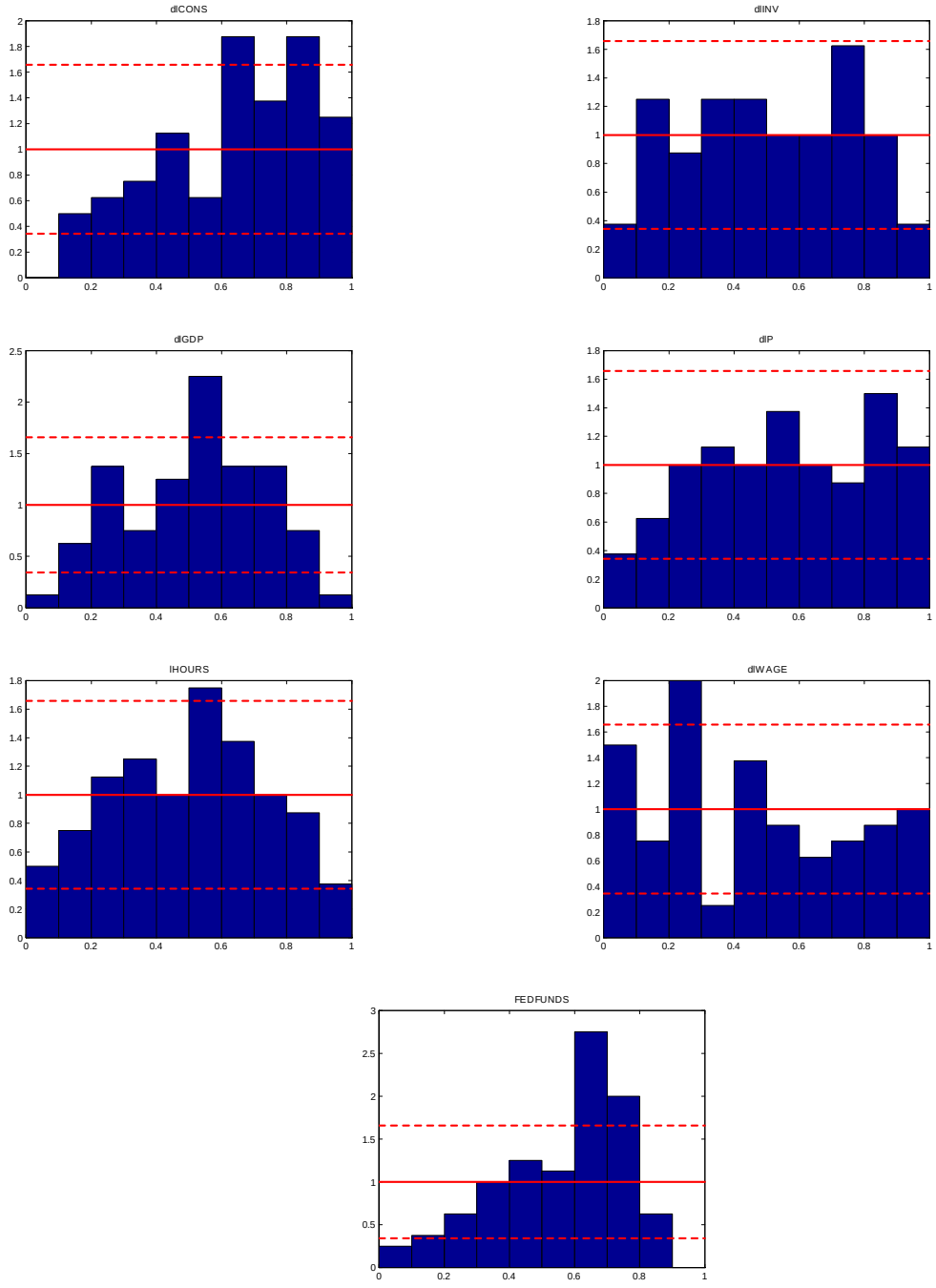






Note: The figures show the empirical CDF of the PITs (solid line), the CDF of the PITs under the null hypothesis (the 45 degree line) and the 95% confidence bands based on critical values of κ_P test reported in Table 1, Panel A. Results are based on a rolling window of size $R = 80$.

Figure 5. PDF of the PITs – Model Forecast Distribution (1985:I-2004:IV)



Note: The figures show the PDF of the PITs (normalized) and the 95% critical values approximated under Diebold et al.'s (1998) binomial distribution (dashed lines), constructed using a normal approximation. The results are based on a rolling window of size $R = 80$.