Alternative Tests for Correct Specification of Conditional Predictive Densities

Barbara Rossi[1] and Tatevik Sekhposyan[2]

January 27, 2016

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 misspecification. An empirical application to the Survey of Professional Forecasters and a baseline macroeconomic model shows the usefulness of our methodology.

Keywords: Predictive Density, Probability Integral Transform, Kolmogorov-Smirnov Test, Cramér-von Mises Test, 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, the 2015 CAMP Workshop on Empirical Macroeconomics and the 2015 CIRANO-CIREQ Workshop on Data Revision in Macroeconomic Forecasting and Policy for comments. B. Rossi gratefully acknowledges financial support from the European Research Agency's Marie Curie Grant 303434 and ERC Grant 615608.

[1]ICREA-Universitat 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 the 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 solid 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 it depends on the horizon for interest rates. The dark, dash-dotted line in the figures plots the actual realized value of the target variable. Clearly, forecasts of interest rates are very imprecise (the realization is outside every forecast decile except for one-quarter-ahead horizon), whereas the model predicts output growth more accurately. Evaluating model-based forecast distributions amounts to understanding whether the model's description of uncertainty was inaccurate or whether the realized values were indeed low probability events.

The methodologies that are currently available 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 at models' *estimated parameter values*, which, we argue, is more empirically useful to measure models' actual forecasting ability in finite samples. In other words, we test whether the predictive densities are correctly specified given the parametric model and the estimation technique specified by the researcher. Accordingly, our test does not require an asymptotic correction for parameter estimation error. Furthermore, our null hypothesis is that of correct specification of the density forecast, which, as we clarify in an example, can still hold even if the forecasting model is dynamically misspecified. Thus, even in the presence of dynamic misspecification, we obtain limiting distributions that are nuisance parameter free for one-step-ahead density forecasts. However, we also extend our framework to multiple-step-ahead forecasts, where the asymptotic distribution of our proposed test is not nuisance parameter free.

Our approach where parameter estimation error is maintained under the null hypothesis is inspired by 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 the test statistics for parameter estimation error; and (ii) the asymptotic distribution of the test statistics is nuisance parameter free even when the model is dynamically misspecified.[5] We derive our tests within the class of Kolmogorov-Smirnov and Cramér-von Mises-type tests commonly used in the literature and show that our proposed tests have good size properties in small

[5] Note that (ii) is not unique to cases where parameter estimation error is maintained under the null hypothesis; in fact, it also holds when parameter estimation error is asymptotically irrelevant, or when one uses martingalization techniques, as in Bai (2003).

samples.

There are several existing approaches for testing the correct specification of a parametric density in-sample (e.g. Bai, 2003, Hong and Li, 2005, Corradi and Swanson, 2006a). Our paper focuses instead on the out-of-sample evaluation of predictive densities. The difference between in-sample and out-of-sample evaluation is that a model may fit well in-sample, and yet its out-of-sample forecasts may be poor (for example, if the distribution of the error changes between the in-sample estimation period and the out-of-sample evaluation period, or if the researcher overfitted the relevant distributional parameters). As such, our paper is related to a series of contributions which test whether observed realizations could have been generated by a given predictive 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, Li and Zhao (2007) and the bootstrap introduced by Corradi and Swanson (2006b,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 predictive ability of the regressors. In the approach we propose the main test statistic is the same as Corradi and Swanson's (2006a) one, although the null hypothesis is very different: it evaluates density forecasts at the estimated parameter values (as opposed to their population values). Thus, our approach is complementary to theirs. Furthermore, since the null hypothesis is different, we cannot directly compare our test to theirs. An alternative approach has been recently proposed by González-Rivera and Sun (2014), who use graphical devices to implement a test of correct specification. The proposed methods work when models are dynamically correctly specified;

[6] Hong, Li and Zhao (2007) provide an out-of-sample counterpart of the Hong and Li (2005) in-sample tests, while Corradi and Swanson (2006b) generalize the in-sample test by Corradi and Swanson (2006a) to an out-of-sample framework. See also Bai and Ng (2005) and Bontemps and Meddahi (2012) for in-sample tests of distributional assumptions.

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.

To illustrate the empirical relevance of our proposed tests, we evaluate density forecasts in the Survey of Professional Forecasters (SPF) and the forecast densities produced by a baseline macroeconomic model. It is very interesting to evaluate the SPF density forecasts using our tests because SPF panelists use a combination of estimated models and expert judgment to produce forecast, even though the models are not disclosed. Thus, in the SPF density forecast case, as well as in most central banks' fan charts, there is parameter estimation error, and it is impossible to correct for it: the only feasible approach is to maintain it under the null hypothesis. This example illustrates the empirical usefulness of our test, since this approach is exactly what we follow in our paper. In fact, we find that predictive densities are, in general, misspecified. In addition, we propose ways to improve the calibration of the densities given our test results.

The remainder of the paper is organized as follows. Section 2 introduces the notation and definitions, and Section 3 discusses issues related to the practical applicability of our test as well as our theoretical approach. In Section 4, we provide Monte Carlo evidence on the performance of our tests in small samples. Section 5 analyzes the empirical applications to SPF and DSGE density forecasts, and Section 6 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 $\{Z_t : \Omega \rightarrow R^{k+1}\}_{t=1}^{T}$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$. The observed vector Z_t is partitioned as $Z_t = (y_t, X_t')'$, where $y_t : \Omega \rightarrow R$ is the variable of interest and $X_t : \Omega \rightarrow R^k$ is a vector of predictors. Let $1 \leq h < \infty$.[7] We are interested in the true, but unknown, $h-$step-ahead conditional predictive density for the scalar variable y_{t+h} based on $\mathcal{F}_t = \sigma(Z_1', ..., Z_t')'$, which is the true information set available at time t. We denote this density by $\phi_0\left(.\right)$.[8]

[7] Note that our framework allows nowcast densities, which technically corresponds to $h = 0$. We do not make this explicit in the notation to avoid misleading the reader to thinking that our tests are in-sample.

[8] The true conditional predictive 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 symbols $\phi_0(.)$ and $\phi_t(.)$ to denote generic distributions and not necessarily a normal distribution.

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 indexed from $t-R+1$ to 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 $\phi_{t+h}\left(.\right)$ 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 $\phi_{t+h}\left(.\right)$ 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

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

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

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

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

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

[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 no breaks. As discussed in Giacomini and White (2006), our proposed approach is also valid for other classes of limited memory estimators.

3 Asymptotic Tests of Correct Specification

This section presents our proposed test for the case of one-step-ahead forecasts. In this case, the tests we propose have an asymptotic distribution that is free of nuisance parameters and the critical values can be tabulated. We further generalize the tests to the presence of serial correlation, which applies to the multi-step-ahead density forecasts. In this case the asymptotic distribution is not nuisance parameter free, and we discuss how to calculate the critical values. All the proofs are relegated to Appendix A.

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:

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

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

The κ_P test statistic is the same as the V_{1T} test statistic considered by Corradi and Swanson (2006a) when applied to predictive densities. Note, however, that we derive 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

[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.

distribution of their test reflects parameter estimation error and, therefore, is not nuisance parameter free. Note that we cannot compare our test with Corradi and Swanson (2006a) since they focus on a different null hypothesis where R tends to infinity, while the theory of our test relies on R being finite. In fact, given that the null hypotheses are different, power in our context corresponds to size in theirs; thus comparisons are not informative.

Under our null hypothesis (eq. 3) instead, the limiting distribution of the test statistic is nuisance parameter free. 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 a few examples.

Example 1: As a simple example, consider $y_{t+1} = c_{t+1} + x_t + \varepsilon_{t+1}$, $\varepsilon_{t+1} \sim iid\ N(0,1)$ and $x_t \sim iid\ N(0,1)$, and ε_{t+1}, x_t are independent of each other. We assume for simplicity that the variance of the errors is known. 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+1} 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 $N\left(c_{t+1} + x_t, 1\right)$. 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:

$$c_{t+1} + x_t = \widehat{\beta}_{t,R} x_t,$$

that is,[11]

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

[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, 2000, p. 104).

Thus, the null hypothesis in eq. (3) is not 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) is 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 is estimated based on a finite number of observations.[12]

Suppose, instead, the true data generating process is: $y_{t+1} = c + x_t + \varepsilon_{t+1}$ where $x_t \sim iid\chi_1^2$ and $\varepsilon_{t+1} \sim iidN(0,1)$. Let the researcher estimate 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.

Example 2: Consider $y_{t+1} = \sigma_t \varepsilon_{t+1}$, $\varepsilon_{t+1} \sim iid\ N(0,1)$ and $\sigma_t^2 = \rho_1 \sigma_{t-1}^2 + \rho_{0,t}$. This is a GARCH(1,0) process for y_{t+1}, where the mean is assumed to be known and equal to zero for simplicity. The researcher instead estimates the model: $y_{t+1} = \gamma e_{t+1}$, $e_{t+1} \sim iid\ N(0,1)$, where the coefficient γ is estimated using observations in a window of size R: $\widehat{\gamma}_t = R^{-1} \sum_{j=t-R+1}^{t} y_j^2$. We set $\rho_{0,t}$ such that our null hypothesis (eq. 3) holds. The estimated PIT and the PIT that generated the data are the same if $\sigma_{t+1}^2 = \widehat{\gamma}_t$. Thus, the null hypothesis is imposed by assuming:

$$\sigma_{t+1}^2 = R^{-1} \sum_{j=t-R+1}^{t} y_j^2 = R^{-1} \sum_{j=t-R+1}^{t} \left(\rho_1 \sigma_{j-1}^2 + \rho_{0,j}\right)^2 \varepsilon_j^2,$$

where $\sigma_{t+1}^2 = \rho_1 \sigma_t^2 + \rho_{0,t+1}$ and

$$\rho_{0,t+1} = R^{-1} \sum_{j=t-R+1}^{t} y_j^2 - \rho_1 \sigma_t^2 = R^{-1} \sum_{j=t-R+1}^{t} \left(\rho_1 \sigma_{j-1}^2 + \rho_{0,j}\right)^2 \varepsilon_j^2 - \rho_1 \sigma_t^2.$$

Example 3: As an example of a dynamically misspecified model where the null hypothesis in eq. (3) holds, consider $y_{t+1} = c_{t+1} + \rho y_t + \varepsilon_{t+1}$, $\varepsilon_{t+1} \sim iid\ N(0,1)$. We assume for

[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.

simplicity that the variance of the errors is known. The researcher instead considers a model $y_{t+1} = \beta + e_{t+1}$, $e_{t+1} \sim iid\ N(0,1)$. Moreover, the researcher is estimating the coefficient β using observations in a window of size R. 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 $N\left(\widehat{\beta}_{t,R}, 1\right)$, where $\widehat{\beta}_{t,R} = R^{-1}\sum_{j=t-R+1}^{t} y_j$, whereas the PIT that generated the data is:

$$\int_{-\infty}^{y_{t+1}} \phi_0\left(y|\mathcal{F}_t\right) dy,$$

where $\phi_0\left(y|\mathcal{F}_t\right)$ is $N\left(c_{t+1} + \rho y_t, 1\right)$. Under the assumption that the variance is known, a sufficient condition for the null hypothesis to hold is that the conditional means from the true DGP and the estimated model are the same. More in detail, the null hypothesis is imposed by assuming:

$$c_{t+1} + \rho y_t = \widehat{\beta}_{t,R},$$

that is,

$$c_{t+1} = \left(R^{-1}\sum_{j=t-R+1}^{t} y_j - \rho y_t\right).$$

It is important to note that R is finite; thus, $R^{-1}\sum_{j=t-R+1}^{t} y_j$ remains a mixing process. In summary, in this case even if the forecasting model is misspecified relative to the data generating process, the null hypothesis in eq. (3), which aims to evaluate correct specification of the model and the forecasting technique jointly, holds.

3.1 One-step-ahead Density Forecasts

This sub-section presents results for the case of one-step-ahead forecasts; the next sub-section generalizes the tests to the presence of 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]

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 $\lambda \in (1, 3/2)$;

[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 and we do not allow either $R \to \infty$ or $P/R \to 0$ in our framework.

(ii) $\Phi_0(y_{t+h}|\mathcal{F}_t)$ is 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 and heterogeneous data. The assumption is similar to the assumption used in Theorem 1 in Giacomini and White (2006), which allows for mild non-stationarity induced by changes in distributions over time, yet rules out I(1) processes. Assumption 1(ii) and 1(iii) are similar to Assumption B in Inoue (2001). These assumptions require the PITs, as well as the marginal and joint distributions of the PITs, to be well-defined.[14] Assumption 1(iv) requires the estimation window size to be finite as the total sample size grows. Our assumptions allow for quite general parametric models (including linear and nonlinear models) and estimation methods (including GMM and MLE), as long as the estimation window size is finite and data are mixing. Furthermore, 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").

Correct specification is characterized by Assumption 2(a):

Assumption 2.
$y_{t+h}|\Im_{t-R+1}^{t} \equiv y_{t+h}|\mathcal{F}_t$ *for all* $t = R, ..., T$, *where* $\equiv$ *denotes equality in distribution.*

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+h} is iid $U(0,1)$, $t = R, ..., T$; (ii) $\Psi_P(r)$ weakly converges to the Gaussian process $\Psi(.)$, with mean zero and auto-covariance function* $E[\Psi(r_1)\Psi(r_2)] = [\inf(r_1, r_2) - r_1 r_2]$.

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 distributions of our proposed test statistics have the appealing feature of being nuisance parameter free. Note that the asymptotic distribution can alternatively be viewed as a Brownian Bridge (see Durbin, 1973).

[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.

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

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

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 $\alpha = 0.10,\ 0.05$ and 0.01 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, from eqs. (1) and (2),

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

Thus,

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

This suggests the following implementation: plot the empirical cumulative distribution function of the PIT, eq. (2), together with the cumulative distribution function of the Uniform (0,1) distribution, r (the 45-degree line), and the critical value lines: $r \pm \kappa_\alpha/\sqrt{P}$. Then, the κ_P test rejects if the cumulative distribution function of the PIT is outside the critical value bands.

We consider two ways of simulating the critical values. One approach, which is what we report in Table 1, relies on simulating the limiting distribution of $\Psi_P(r)$, considered in Theorem 1, directly. More specifically,

(i) Discretize the grid for r. In particular, we consider the grid: $r = [0 : 0.001 : 1]$;

[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 and density forecast comparisons.

(ii) Calculate the theoretical variance $E[\Psi(r_1)\Psi(r_2)] = [\inf(r_1, r_2) - r_1 r_2]$, for r_1, $r_2 \in [0 : 0.001 : 1]$;

(iii) Draw independent multivariate Normal random variables based on the Cholesky decomposition of the estimated covariance matrix $E[\Psi(r_1)\Psi(r_2)]$ calculated in (ii);[16]

(iv) Construct the test statistics proposed in Theorem 2**;**

(v) Repeat the steps (iii) and (iv) for a large number of Monte Carlo replications; report the $90\%, 95\%$ and 99% percentiles of the simulated limiting distribution as critical values.

The second approach aims at obtaining exact critical values in finite samples. We do so by the following procedure:

(i) Draw P independent random variables from the Uniform (0,1) distribution;

(ii) For a given sample of size P, construct the $\Psi_P(r)$ as in eq. (1);

(iii) Construct the test statistics proposed in Theorem 2;

(iv) Repeat steps (i) to (iii) for a large number of Monte Carlo replications; the 90%, 95% and 99% percentile values of the simulated limiting distribution are the critical values.

The second approach has the advantage that we can tailor the critical values to the sample sizes used in an empirical application. However, it has the disadvantage that it is computationally slower, as it involves the calculation of a sum $(\Psi_P(r))$. As we show later in a Monte Carlo size experiment (presented in Table 2), the finite-sample critical values improve the size, more so for smaller values of P.

In principle, the limiting distribution of the Kolmogorov-Smirnov test reported in Theorem 1 is not different from that of the usual textbook version of the Kolmogorov-Smirnov test that are derived by analytical calculations (see Durbin, 1973). According to Smirnov (1948) these values are 1.22, 1.36 and 1.63 for $\alpha = 0.10, 0.05$ and 0.01, respectively. Unreported Monte Carlo size experiments suggest that our simulated critical values result in tests that have better size than those based on the theoretical formula: the theoretical formula appears to be more conservative than our simulation-based procedure.

It is interesting to compare our approach to Diebold et al. (1998). While our null hypothesis is different from theirs, the procedure that we propose is similar to theirs in that both their implementation and ours abstract from parameter estimation error (for different reasons). 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 critical value bands that we propose are joint, not pointwise.

[16] A finer grid, in theory, could results in a more accurate approximation, but numerically the Cholesky factorization becomes less accurate. Thus, there is a limited payoff from refining the grid.

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 density forecast is correctly specified, an inverse Normal transformation of the PITs (ζ_{t+h}) has a standard Normal distribution (Berkowitz, 2001).[17] 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 Berkowitz's (2001) likelihood ratio tests to account for parameter estimation error.

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

Thus, one could test for the correct specification of the density forecast by testing the absence of serial correlation and the correct specification of specific moments of ζ_{t+1}. 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. This approach has the advantage of being 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 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. This is a more general case, and we consider it in the next section.

3.2 Multi-step-ahead Forecasts

When considering h-step-ahead forecasts, $h > 1$ and finite, an additional problem arises as the PITs become serially correlated. Thus, we need to extend our results and allow the

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

forecasts to be serially correlated under the null hypothesis; that is, when Assumption 2 does not hold.

When evaluating h-step-ahead conditional predictive densities, the next Theorem shows that $\Psi_P(r)$ weakly converges 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} is $U(0,1)$, $t = R, ..., T$; (ii) $\Psi_P(r)$ weakly converges 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,*

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

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

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. In the Monte Carlo we obtain the critical values of our tests using the Newey and West's (1987) HAC estimator for the $(r \times r)$ dimensional covariance of $\xi_{t+h}(r)$. While HAC estimates can be sensitive to the choice of the bandwidth, for multi-step ahead forecasts the serial correlation is of order $(h-1)$ under H_0, which provide some guidance on the choice of the bandwidth.

More specifically, we calculate the critical values based on the method described below Theorem 2 with a notable difference. Namely, when simulating the critical values directly based on the asymptotic multivariate Normal distribution, we use the Cholesky factorization of $E[\Psi(r_1)\Psi(r_2)] = \sigma(r_1, r_2)$, which we estimate with the Newey and West (1987) HAC on the estimated $\xi_{t+h}(r)$. Appendix A discusses the HAC estimator.

In this case, the limiting distribution resembles the one that Corradi and Swanson (2006c) obtain under dynamic misspecification since the limiting distribution is not free from parameter estimation error. However, under the null hypothesis, we are not concerned about dynamic misspecification since the null hypothesis may hold even though a model can be dynamically misspecified (see Example 3 in the previous section).

There are several other solutions proposed in the literature that one could use within our approach as well. One 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 by creating sub-samples of predictive distributions that are 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 Monte Carlo Evidence

In this section we analyze the size and power properties of our proposed tests in small samples for both one- and multi-step-ahead 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.

4.1 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.[18] 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

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

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

[18] Note that Corradi and Swanson (2006b) focus on a different null hypothesis based on $R \to \infty$, and the theory of our test instead relies on R being finite. In fact, given that the null hypotheses are different, power in our context corresponds to size in their context; thus comparisons are meaningless.

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: $y_t = \mu_t + \gamma x_{t-1} + \varepsilon_t$, $\varepsilon_t \sim iidN(0,1)$, $x_t = 0.2 + 0.8x_{t-1} + \nu_t$, $\nu_t \sim iid\ N(0, 1.08^2)$ and is independent from ε_t, $\gamma = 0.48$ and

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

DGP S3 (GARCH): Consider the data being generated by a GARCH(1,0), where $y_t = \sigma_t \varepsilon_t$, $\varepsilon_t \sim iidN(0,1)$ and the $\sigma_t^2 = \rho_{0,t} + \rho_1 \sigma_{t-1}^2$. On the other hand, the forecasting model is $y_t = \gamma_t e_t$, $e_t \sim iidN(0,1)$, and

$$\rho_{0,t} = \left(R^{-1} \sum_{j=t-R}^{t-1} \left(\rho_{0,j} + \rho_1 \sigma_{t-1}^2 \right) \varepsilon_j^2 - \rho_1 \sigma_{t-1} \right).$$

DGPs S1-S3 are based on one-step-ahead forecast densities. DGP S4 considers the case of h-step-ahead forecast densities ($h = 2$) where the PITs are serially correlated by construction.

DGP S4 (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 for DGP S1 are shown in Table 2. The table shows that our proposed tests have good size properties. Further, the finite sample critical values improve the performance of the test for small values of P (see Panel B). Table 3 shows that our tests perform well in finite samples in DGPs S2-S4, with under-rejections in DGP S2 for large values of R. In the case of serial correlation, DGP S4, the asymptotic distribution of the tests in Theorem 3 approximated using HAC-consistent variance estimates tends to slightly over-reject, particularly for small values of R.[19]

INSERT TABLES 2 AND 3 HERE

4.2 Power Analysis

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

[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).

DGP P1: The data are generated from a linear combination of Normal and χ_1^2 distributions: $y_t = \mu_t + x_{t-1} + (1-c)\,\widehat{\sigma}_t \eta_{1,t} + c\left(\eta_{2,t}^2 - 1\right)\sqrt{2}$, 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 density becomes a convolution of a standard Normal and a χ_1^2 distribution (with mean zero and variance one), where the weight on the latter becomes larger as c increases.[20]

DGP P2: We estimate a model $y_t = \beta x_{t-1} + e_t$, $e_t \sim iidN(0,1)$. The data is generated by $y_t = \mu_t + x_{t-1} + \varepsilon_t$, $\varepsilon_t \sim iid\ t_\nu$, where $x_t \sim iidN(0,1)$, μ_t is defined as in DGP S1**,** while ν is the number of degrees of freedom. When ν is large, the null hypothesis is satisfied; as ν decreases, the misspecification increases.

The results shown in Table 4 suggest that our proposed specification tests (κ_P, C_P) have good power properties in detecting misspecification in the predictive density.[21]

INSERT TABLE 4 HERE

5 Empirical Analysis

This section provides an empirical assessment of the correct specification of widely-used density forecasts: (i) the Survey of Professional Forecasters (SPF) density forecasts of inflation and output growth; and (ii) density forecasts of the seven macroeconomic aggregates in a representative macroeconomic model. The reasons why we focus on these two examples are as follows. First, we focus on the SPF because SPF panelists use a combination of estimated models and expert judgment to produce forecast, even though the models are not known and, even if they were, they apply expert judgment to trim the forecasts and/or combine models' forecasts. In fact, in a recent SPF survey overview, Stark (2013, p. 2) found that: "Overwhelmingly, the panelists reported using mathematical models to form their projections. However, we also found that the panelists apply subjective adjustments to their pure-model forecasts. The relative role of mathematical models changes with the forecast horizon." Interestingly, the survey also found that SPF panelists "change their forecasting approach with the length of the forecast horizon. At the shortest horizons (two years

[20] Note that $\left(\eta_{2,t}^2 - 1\right)\sqrt{2}$ 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. $T = 100$.

out and less), mathematical models are widely used by the panelists. Between 18 to 20 forecasters reported using models at these short horizons (...). Panelists also reported using models for long-horizon projections as well (three or more years out), although proportionately fewer rely on models at the long horizons than at the short horizons. (...) They use a combination of models in forming their expectations, rather than just one model." Thus, in the SPF density forecast case, it is impossible to correct for parameter estimation error: the only approach is to maintain it under the null hypothesis. This is exactly the approach we follow in our paper. This highlights the empirical usefulness of the methodologies described in our paper. In fact, one of the advantages of our testing approach is that the only information needed for the implementation is a predictive density: knowledge of the model that generated the forecasts is not necessary.

Second, we evaluate forecast densities from a DSGE model. Note that the forecast density to be evaluated in our framework can be obtained in many ways, either frequentist or Bayesian. In fact, in this paper, we are proposing a general way to evaluate forecast distributions; in particular, if one is interested in evaluating whether a forecast distribution obtained by any method (including Bayesian methods) is correctly specified using frequentist methods, one can use the method we propose. There are several cases in the literature where a model is estimated with Bayesian methods and yet inference is based on PITs, e.g. Clark (2011).

5.1 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

[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.

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.

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. To evaluate the density forecasts we use the year-over-year growth rates of output and prices calculated from the first quarterly vintage of the real GNP/GDP and the GNP/GDP deflator in levels. 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 5. Asterisks ('*') indicate rejection at the 5% significance level based on the critical values in Theorem 2 (reported in Table 1, Panel A). The test rejects correct specification for both output growth and inflation, except for output growth at the one-year-ahead forecast horizon.

INSERT TABLE 5 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, in contrast, survey forecasts do not characterize the predictive distribution of inflation correctly.

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 bands, 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. For instance, in the case of the current year output growth forecasts in Panel A of Figure 2, it appears that there are not as many realization in the left tail of the distribution relative to what the forecasters expected (the slope in the left tail is flat relative to the 45 degree line). In the case of the current year inflation forecast in Panel B, the forecasters overestimate both tails of the distribution and, instead, do not put as much probability on potential outcomes in the middle of the distribution. One-year-ahead inflation forecasts appear to have a bit of a different dynamics: there is no evidence of misspecification in the left tail, but that are many more realization relative to expected frequencies in the left half of the distribution. This also comes at the expense of misspecifying the right tail of the distribution: the forecasters are more optimistic about extremely positive output growth scenarios relative to the realized outcomes.

INSERT FIGURES 2 AND 3 HERE

For comparison, Figure 3 reports results based on Diebold et al.'s (1998) test. Panel A plots the histogram 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 depict the 95% confidence interval (dotted lines) using a Normal approximation to a binomial distribution similar to Diebold et al.'s (1998). Both current year and one-year-ahead density forecasts of output growth in Panel A are misspecified, although misspecification is milder in the case of one-year-ahead output growth. Figure 3, Panel B, shows the histogram of 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 assuming iid-ness of the PITs. The discrepancy in the results is most likely due to the fact that the latter test is pointwise (for each bin), whereas we jointly test the correct specification across all quantiles in the empirical distribution function: thus, in order to correctly account for the joint null hypothesis, our test has larger critical values than theirs.

Once our tests reject, it is of interest to investigate how one can improve the calibration of the density forecast. Consider, for example, the SPF's predictive densities of inflation. Figure 4 plots the historical evolution of the mean probability forecasts of inflation (solid line), together with the 2.5-th, 5-th, 50-th, 95-th and 97.5-th percentiles of the predictive distribution. The picture shows that the density forecasts for both the current year and the next have evolved over time; in particular, they have become tighter towards the end of the sample. When compared to the realization (dash-dotted line), the forecast distribution looks reasonable, as the realization is contained within the 90% confidence interval throughout the sample period.

INSERT FIGURES 4 AND 5 HERE

However, the visual evidence is misleading: after studying the PITs and implementing our test (whose results are depicted in Figures 2 and 3, Panel B), we conclude that the distribution is not correctly calibrated, i.e. on average the realizations are not consistent with the ex-ante probabilities of potential outcomes. Moreover, for the case of one-year-ahead forecasts (right-hand figure in Panel B) our test finds that there are many more realizations below the mean relative to what has been anticipated. A careful observation of Figure 4 would reinforce that evidence: frequently, the realization of one-year-ahead inflation is below the forecasted mean of the distribution. Thus, the pictures suggest that the distribution is misspecified and the reason is the misspecification of the mean. To investigate this formally, one can test for forecast unbiasedness; that is, test whether $\alpha = 0$ in the regression: $y_{t+1} - \widehat{y}_{t+1} = \alpha + \varepsilon_t$. The full sample estimate is $\widehat{\alpha} = -0.69$ with a t-statistics of -5.38.[23] The results appear to be consistent with the message in the figures.

In fact, after adjusting the mean of the distribution by adding the estimated bias (depicted in Figure 5, left panel), one obtains a well-calibrated distribution: the right panel in Figure 5 shows the results of the test for correct calibration after the (infeasible) bias adjustment, and confirms that indeed the correct specification is not rejected by our test.

To summarize, this example shows that our test can be used as a first step to determine whether the forecast density is correctly calibrated; if our test rejects the correct calibration of the forecast density, an additional analysis of the plot of the test statistic can provide guidance on the possible sources of the problem; additional (forecast rationality) tests can

[23] The t-statistics is constructed with a Newey-West HAC estimator for the variance. If one were worried about presence of instabilities, one could alternatively apply an unbiasedness test robust to instabilities – see Rossi and Sekhposyan (2015).

then verify the conjecture and help improve the calibration of the density forecast for the future (provided the source of the misspecification does not change over time).

5.2 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 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 forecast densities 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 forecast densities 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 windows of 80 observations and produce a sequence of 80 out-of-sample forecast densities.[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 6 reports the empirical results for the correct calibration of the model's forecast densities. 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). According to the critical values in Theorem 2, the density forecasts of investment, inflation, hours and wages are correctly calibrated. On the other hand, for consumption, output growth and the federal funds rate, we reject the null hypothesis of correct calibration.

INSERT TABLE 6 HERE

Figure 6 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 and output growth in the lowest quantiles of the distribution; that is, the model over-predicts the lower tail values of the target variable. For the federal funds rate the misspecification is in both tails: there

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

[25] Dynare (http://www.dynare.org), the software package used by Smets and Wouters (2007), approximates the deciles of the forecast distribution with a Gaussian kernel, given the DSGE's assumption of normally distributed errors. We obtain the PITs using a linear interpolation for the inter-decile range. Alternatively, one could focus on the Bayesian predictive density. However, the forecast distribution is the typical outcome of Dynare's estimation routine, thus we focus on it.

[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.

are fewer realizations in either tail relative to the frequency by which they have been anticipated. For the remaining variables, the test suggests proper calibration. For comparison, Figure 7 shows the histograms 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 (per bin).

INSERT FIGURES 6 AND 7 HERE

6 Conclusions

This paper proposes new tests for predictive density evaluation. The techniques are based on Kolmogorov-Smirnov and Cramér-von Mises-type test statistics and focus both on the whole distribution as well as specific parts of it. We also propose methodologies that can be applied to multiple-step-ahead forecast horizons. Our empirical analyses uncover that both SPF output growth and inflation density forecasts as well as DSGE-based forecasts densities of several macroeconomic aggregates are misspecified. We also investigate possible avenues that practitioners may follow in order to improve the density forecasts using our test results.

[27] Note that this is a fair comparison, since both Figures 6 and 7 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] Christoffel, K., G. Coenen and A. Warne (2010), "Forecasting with DSGE Models," *ECB Working paper* 1185.

[11] Clark, T. (2011), "Real-Time Density Forecasts from VARs with Stochastic Volatility," *Journal of Business and Economic Statistics* 29(3), 327-341.

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

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

[14] 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.

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

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

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

[18] 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.

[19] 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.

[20] 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.

[21] Durbin, J. (1973), *Distribution Theory for Tests Based on the Sample Distribution Function*. Philadelphia: SIAM.

[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 (2014), *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] Rossi, B. and T. Sekhposyan (2015), "Forecast Rationality Tests in the Presence of Instabilities, With Applications to Federal Reserve and Survey Forecasts", *Journal of Applied Econometrics*, forthcoming.

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

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

[44] Smirnov, N . (1948), "Table for Estimating the Goodness of Fit of Empirical Distributions," *Annals of Mathematical Statistics* 19, 279-281.

[45] Stark, T. (2013), "SPF Panelists' Forecasting Methods: A Note on the Aggregate Results of a November 2009 Special Survey," *Real-Time Data Research Center Research Discussion Paper*.

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

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

[48] 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 3. The sequence of the proofs is as follows. First, we prove Theorem 4, then proceed to proving Theorem 1, which follows from Theorem 4 under Assumption 2; finally, we prove Theorems 2 and Corollary 3.

Proof of Theorem 4. (i) Under Assumption 1(ii) and H_0 in eq. (3), by the proof of 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 (2006), 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) $r = 1$. Note that the limiting variance can be consistently estimated both under the null and the alternative following Inoue (2001, p. 165) under the additional assumptions in Inoue (2001). ■

Proof of Theorem 1. (i) Follows from Bai (2003, lemma 1).

(ii) 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 i.i.d. Uniform(0,1). ■

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). ■

Appendix B. Tables and Figures

Table 1. Critical Values

		κ_α			C_α		
	α :	0.01	0.05	0.10	0.01	0.05	0.10
Panel A. Tests on the Whole Distribution							
Correct Specification Test		1.607	1.339	1.205	0.742	0.460	0.347
Panel B. Tests on Specific Parts of the Distribution							
Left Tail	$r \in [0, 0.25]$	1.240	0.996	0.876	0.559	0.335	0.243
Left Half	$r \in [0, 0.50]$	1.535	1.255	1.115	0.856	0.516	0.377
Right Half	$r \in [0.50, 1]$	1.532	1.254	1.116	0.854	0.515	0.378
Right Tail	$r \in [0.75, 1]$	1.238	0.997	0.876	0.562	0.336	0.243
Center	$r \in [0.25, 0.75]$	1.605	1.330	1.191	1.183	0.714	0.523
Tails	$r \in \{[0, 0.25] \cup [0.75, 1]\}$	1.327	1.100	0.987	0.412	0.270	0.212

Note: Panel A reports the 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 the critical values for the same statistics for specific parts of the distributions, indicated in the first and second columns. The number of Monte Carlo replications is 1,000,000. The domain for r is discretized with $r =[0 : 0.001 : 1]$.

Table 2: Size Properties

Panel A: DGP S1 (Asymptotic Critical Values)									
		κ_P				C_P			
P	$R:$	25	50	100	200	25	50	100	200
25		0.044	0.042	0.044	0.041	0.047	0.048	0.049	0.046
50		0.053	0.052	0.052	0.053	0.053	0.052	0.055	0.053
100		0.047	0.045	0.050	0.050	0.048	0.047	0.045	0.053
200		0.053	0.050	0.052	0.054	0.050	0.051	0.054	0.050
500		0.050	0.054	0.060	0.052	0.049	0.053	0.051	0.051
1000		0.050	0.046	0.044	0.053	0.051	0.048	0.049	0.055

Panel B: DGP S1 (Finite Sample Critical Values)									
		κ_P				C_P			
P	$R:$	25	50	100	200	25	50	100	200
25		0.047	0.046	0.047	0.048	0.047	0.048	0.045	0.047
50		0.054	0.055	0.052	0.057	0.055	0.048	0.054	0.056
100		0.044	0.046	0.051	0.049	0.048	0.046	0.045	0.049
200		0.053	0.053	0.060	0.049	0.052	0.056	0.058	0.046
500		0.044	0.051	0.055	0.056	0.045	0.053	0.050	0.053
1000		0.049	0.047	0.051	0.056	0.052	0.049	0.051	0.055

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, $r = [0 : 0.001 : 1]$. The critical values used for Panel A are based on the asymptotic distribution, reported in Table 1, Panel A, while the critical values used for Panel B are based on simulated finite-sample distributions with 5,000 Monte Carlo replications.

Table 3: Size Properties

DGP S2 (IID Case)									
		κ_P				C_P			
P	$R:$	25	50	100	200	25	50	100	200
25		0.062	0.047	0.042	0.042	0.067	0.053	0.050	0.046
50		0.062	0.049	0.042	0.047	0.064	0.048	0.047	0.054
100		0.060	0.033	0.037	0.045	0.058	0.032	0.040	0.043
200		0.055	0.036	0.033	0.043	0.056	0.030	0.027	0.040
500		0.061	0.030	0.032	0.031	0.057	0.026	0.026	0.023
1000		0.049	0.031	0.023	0.021	0.051	0.025	0.021	0.022
DGP S3 (GARCH Case)									
		κ_P				C_P			
P	$R:$	25	50	100	200	25	50	100	200
25		0.043	0.040	0.049	0.045	0.045	0.042	0.053	0.049
50		0.048	0.046	0.043	0.047	0.050	0.049	0.047	0.051
100		0.047	0.051	0.049	0.047	0.050	0.055	0.047	0.050
200		0.050	0.053	0.050	0.051	0.050	0.055	0.050	0.054
500		0.056	0.059	0.049	0.050	0.052	0.051	0.047	0.048
1000		0.049	0.050	0.046	0.053	0.050	0.051	0.048	0.056
DGP S4 (Serially Correlated Case)									
		κ_P				C_P			
P	$R:$	25	50	100	200	25	50	100	200
25		0.096	0.076	0.046	0.103	0.135	0.061	0.056	0.070
50		0.077	0.092	0.121	0.123	0.055	0.087	0.117	0.114
100		0.072	0.102	0.069	0.056	0.048	0.075	0.069	0.078
200		0.061	0.082	0.089	0.072	0.051	0.059	0.086	0.079
500		0.051	0.082	0.069	0.068	0.052	0.075	0.071	0.076
1000		0.074	0.072	0.068	0.072	0.072	0.067	0.074	0.066

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, $r = [0 : 0.001 : 1]$. Critical values used for DGPs S2-S3 are in Table 1, Panel A. For DGP S4, they are simulated based on 10,000 Monte Carlo replications with a HAC covariance matrix estimate.

Table 4. Power Properties

DGP P1			DGP P2		
c	κ_P	C_P	ν	κ_P	C_P
0	0.061	0.062	30	0.051	0.057
0.15	0.065	0.066	10	0.068	0.057
0.30	0.080	0.088	7	0.154	0.100
0.35	0.131	0.131	6	0.253	0.181
0.40	0.252	0.244	5	0.485	0.439
0.45	0.516	0.505	4	0.855	0.874
0.50	0.854	0.872	3	0.999	0.999
0.60	1.000	1.000	2	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, $r = [0 : 0.001 : 1]$. We use the critical values reported in Table 1, Panel A, to calculate the empirical rejection frequencies.

Table 5: Correct Specification Tests for SPF's Probability Forecasts

Series Name:	GDP Growth		GDP Deflator Growth	
	κ_P	C_P	κ_P	C_P
Current-year Forecasts	1.534*	0.773*	3.979*	5.066*
One-year-ahead Forecasts	0.821	0.110	3.734*	5.890*

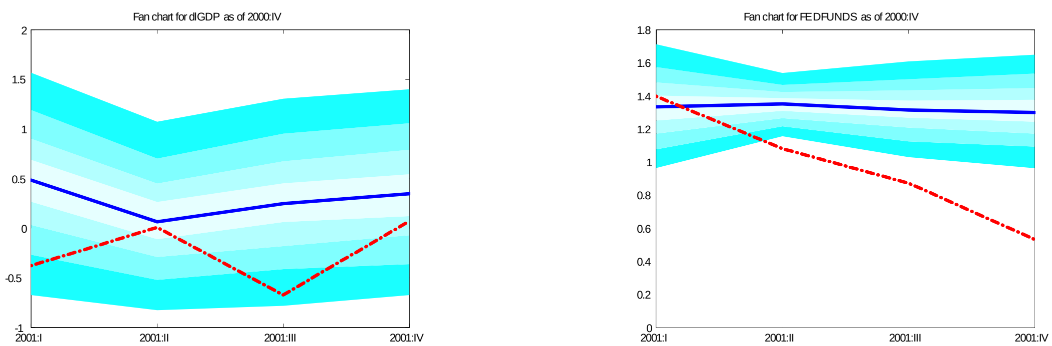
Note: Asterisks '*' indicate rejection at 5% significance level based on the critical values in Theorem 2 (reported in Table 1, Panel A). The domain for r is discretized, $r = [0 : 0.001 : 1]$.

Table 6: Correct Specification Tests for Model Forecast Distribution

Variable	κ_P	C_P
Consumption (real)	2.218*	1.918*
Investment (real)	0.693	0.150
Output Growth (real)	1.440*	0.662*
Inflation	1.239	0.456
Hours	0.966	0.301
Wages (real)	1.145	0.254
Federal Funds Rate	1.874*	1.380*

Note: 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). The domain for r is discretized, $r = [0 : 0.001 : 1]$. The evaluation sample is from 1985:I - 2004:IV.

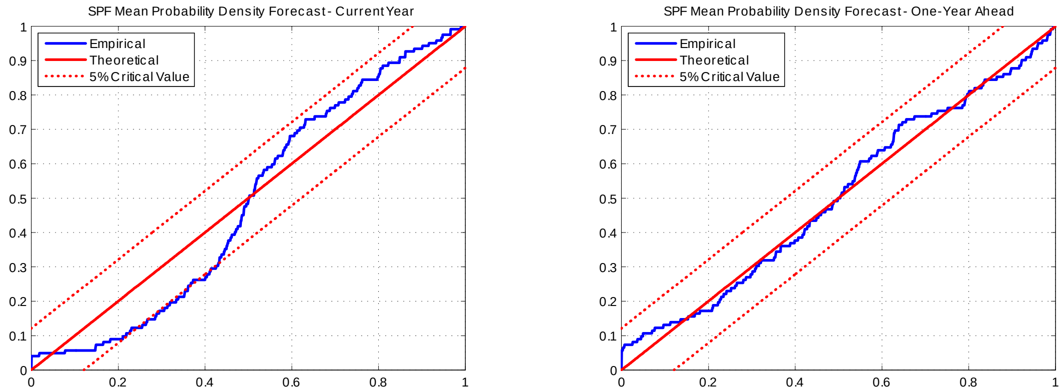
Figure 1. Representative Fan Charts from a Representative Model in 2000:IV



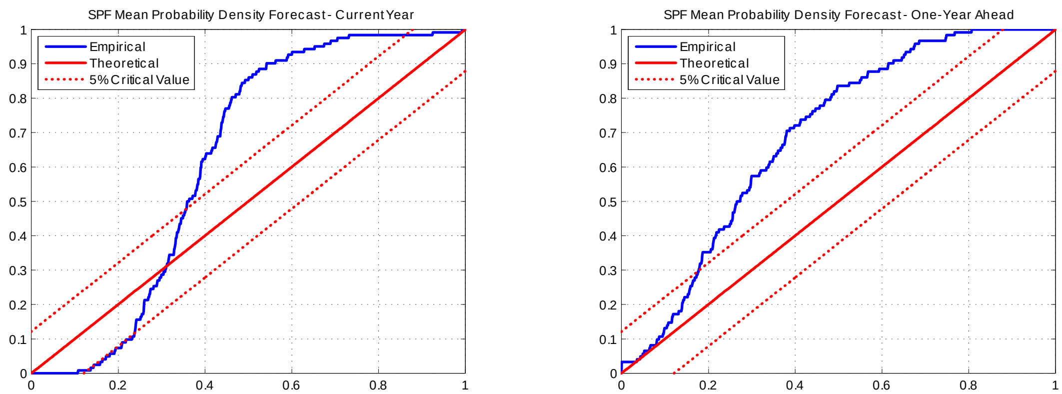
Note: The figure shows fan charts obtained by estimating a baseline model with data up to 2000:IV, prior to the 2001:I-2001:IV recession. The shaded areas depict the 10-th, 20-th, 30-th, 40-th, 50-th, 60-th, 70-th, 80-th and 90-th deciles of the forecast distribution for one to four-quarter-ahead out-of-sample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

Figure 2. CDF of the PITs – SPF Probability Forecast

Panel A: GDP Growth



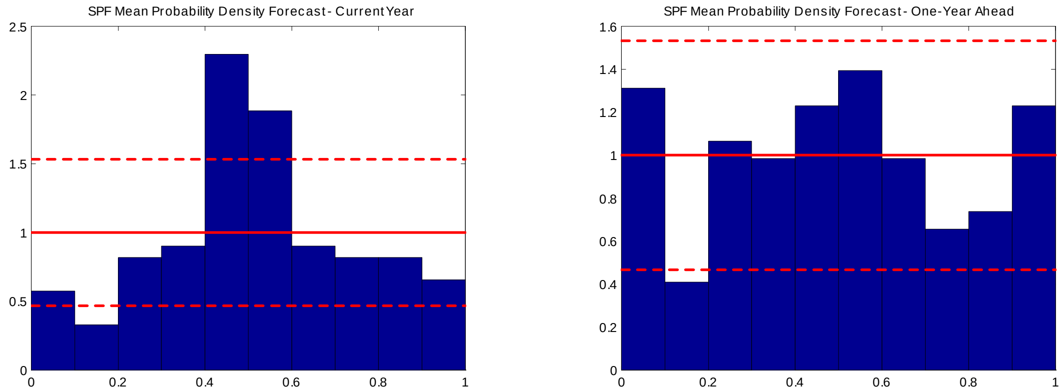
Panel B: GDP Deflator Growth



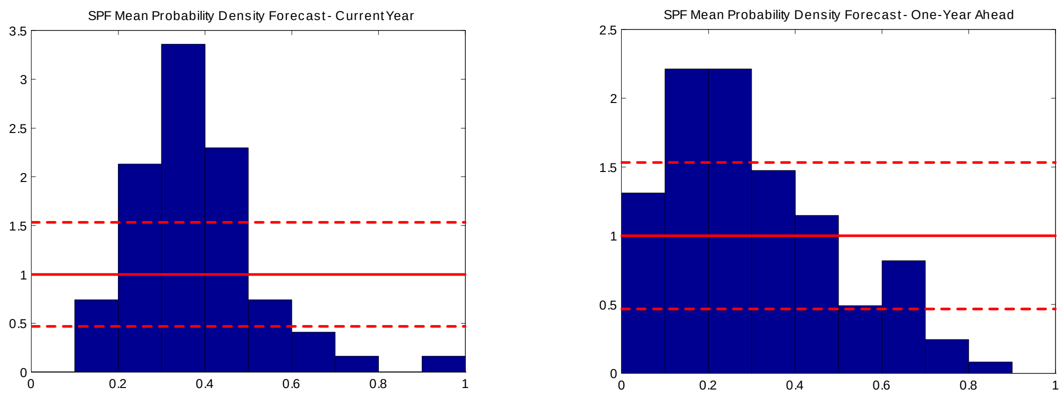
Note: The figure shows the empirical CDF of the PITs (solid line), the CDF of the PITs under the null hypothesis (the 45 degree line) and the 5% critical values bands based on the κ_P test reported in Table 1, Panel A.

Figure 3. Histogram of the PITs – SPF Probability Forecast

Panel A: GDP Growth

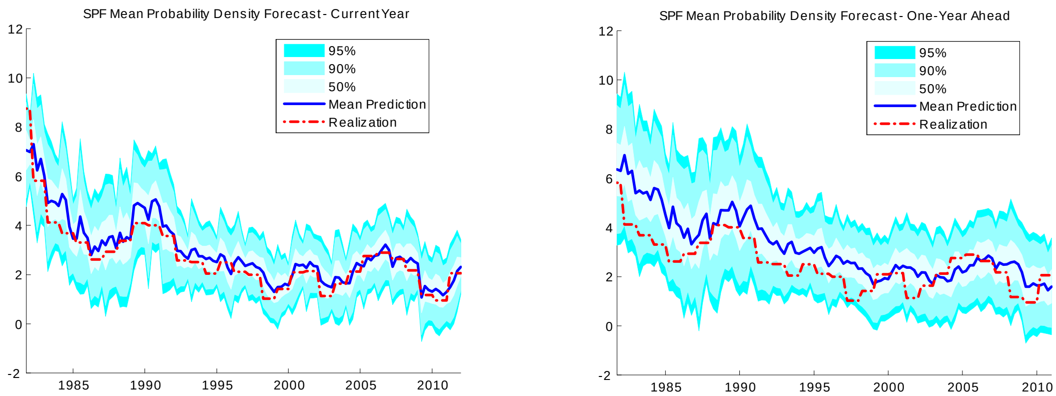


Panel B: GDP Deflator Growth



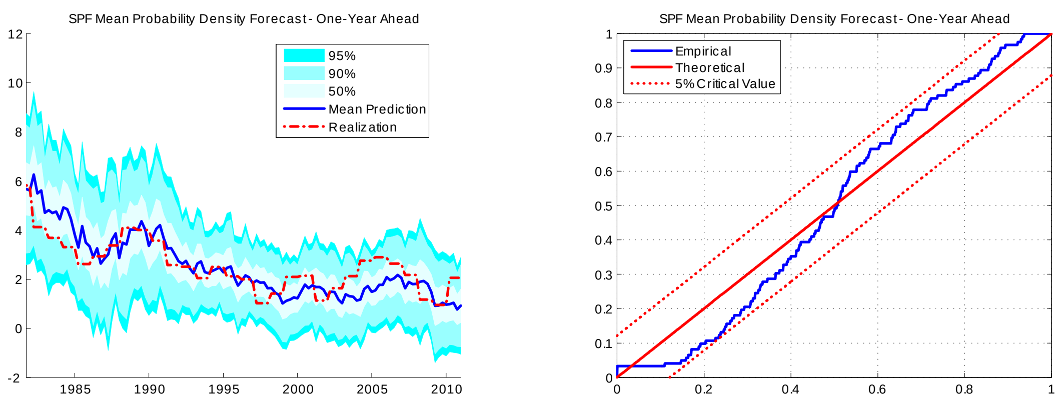
Note: The figures show the histograms of the PITs (normalized) and the 95% confidence interval approximated by Diebold et al.'s (1998) binomial distribution (dashed lines), constructed using a Normal approximation.

Figure 4. Mean and Quantiles of the SPF Inflation Density Forecast



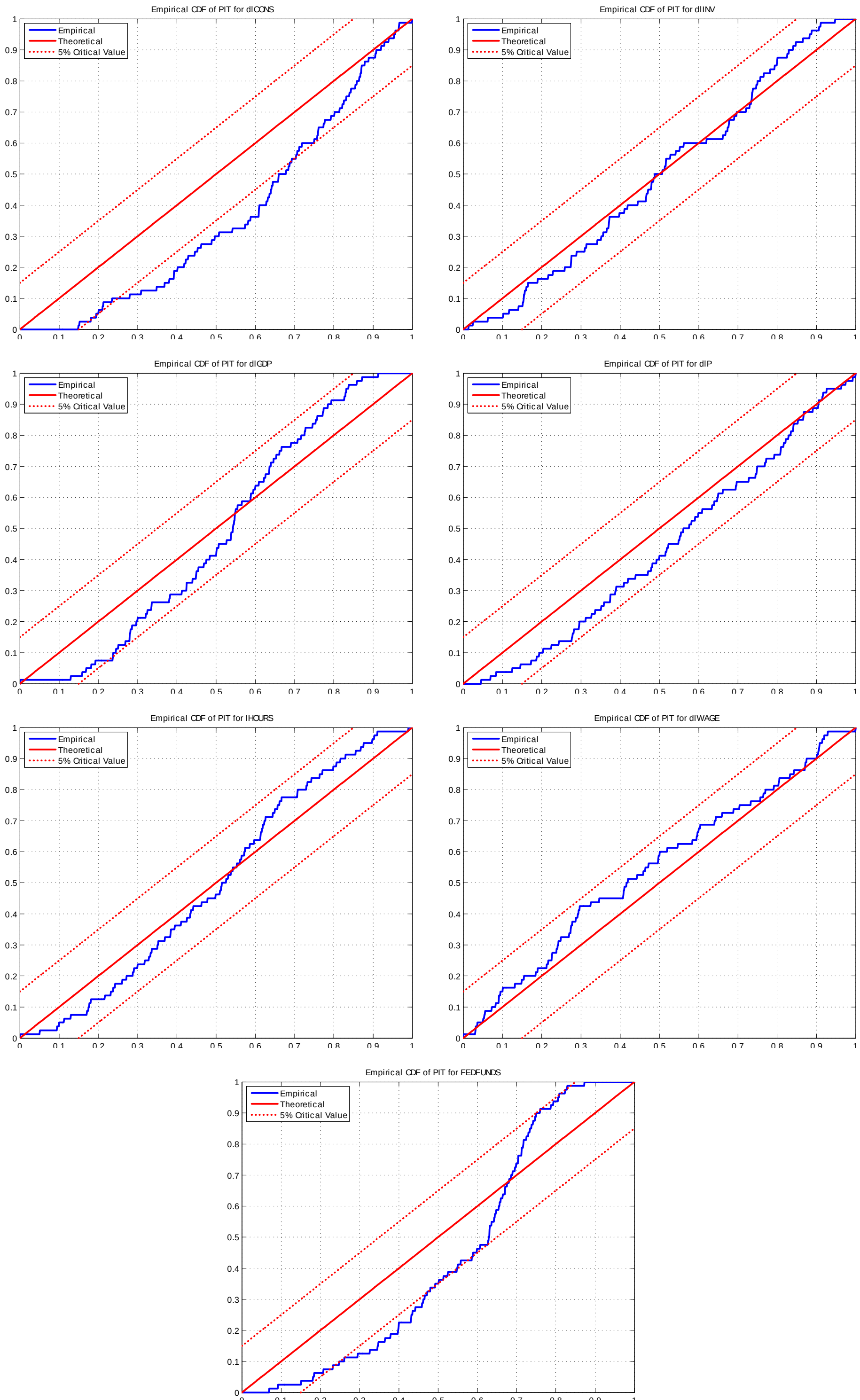
Note: The figures plot quantiles of the SPF density forecast over time. The quantiles are constructed based on a normality assumption on the average SPF density forecasts at each point in time.

Figure 5: Forecast Evaluation of Bias-adjusted SPF Inflation



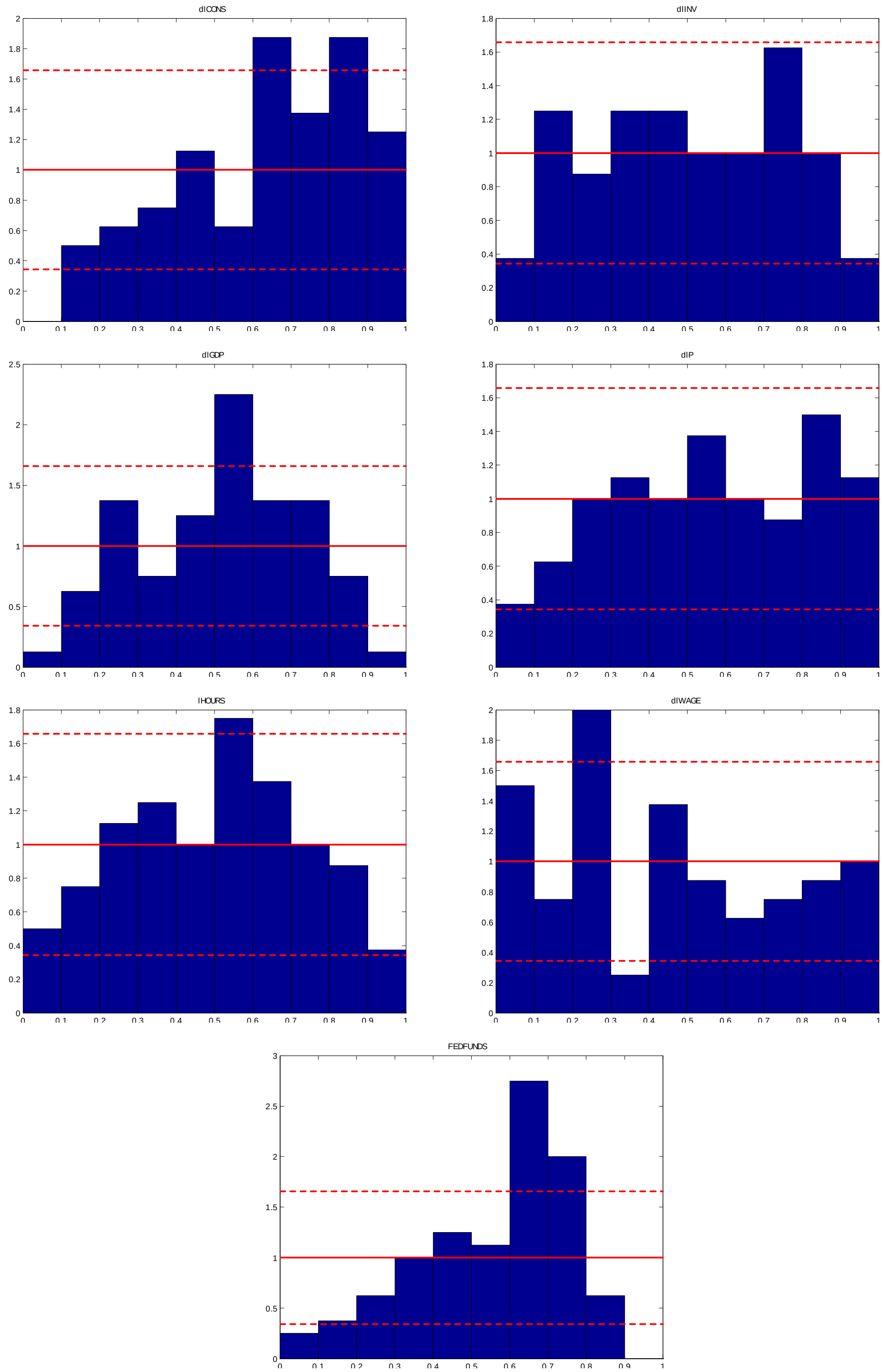
Notes: The figures show quantiles of the SPF density forecast of next year's inflation (left panel), as well as the test of correct calibration (right panel) after the correction to account for the average bias of the SPF in the observed sample.

Figure 6. CDF of the PITs – Model Forecast Distribution



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 5% critical value bands reported in Table 1, Panel A. Estimation window size is $R = 80$.

Figure 7. Histogram of the PITs – Model Forecast Distribution



Note: The figures show the histograms of the PITs (normalized) and the 95% confidence interval approximated under Diebold et al.'s (1998) binomial distribution (dashed lines), constructed using a Normal approximation. Estimation window size is $R = 80$.