Tests of Specification and Distributional Change for Predictive Densities

Barbara Rossi* and Tatevik Sekhposyan†

March 1, 2012

Abstract

We propose new methods for evaluating predictive densities. First, we propose new Kolmogorov-Smirnov and Cramér-von Mises-type tests for correct specification of predictive densities that are robust to dynamic misspecification and to the presence of instabilities. In addition, we offer a simple way of testing for distributional change in predictive densities even if they are mis-specified. Our results indicate that our tests are well sized and have good power in detecting misspecification and time variation (individually and jointly) in predictive densities. An empirical application to the density forecasts of the Survey of Professional Forecasters shows the usefulness of our methodology.

Keywords: Predictive Density, Dynamic Misspecification, Instability, Structural Change, Forecast Evaluation

Acknowledgments: We thank T. Clark, F. Diebold, A. Inoue, N. Swanson, participants of the seminars at the Bank of Canada, Chicago Booth, UNC Chapel Hill and the 2012 American Economic Association meetings for comments. The views expressed in this paper are those of the authors. No responsibility should be attributed to the Bank of Canada.

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

*ICREA, Pompeu Fabra U., CREI and Barcelona Graduate School of Economics; and Duke University. E-mail: barbara.rossi@upf.edu

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

1 Introduction

Predictive densities provide a measure of uncertainty around mean forecasts, thus enabling researchers to quantify the risk in forecast-based decisions. For example, they are useful tools for central banks and policymakers, as they enable them to take into account forecast uncertainty in their economic decisions. It is therefore important to have tools to evaluate whether predictive densities are correctly specified. Diebold et al. (1998, 1999) introduced the probability integral transform (PIT, Rosenblatt, 1952) to economics and finance as a tool to test whether a predictive distribution matches that of the true (and unobserved) distribution that generates the data. Subsequent contributions extended their PIT methodology to account for parameter estimation error and dynamic misspecification. The former requires an adjustment to the PIT to account for the uncertainty associated with parameter estimation. Dynamic misspecification implies that the information available to a researcher spans only a subset of the information the true model is in fact conditioned on. Among recent contributions, Bai (2003) and Hong and Li (2003) propose tests for correct specification aimed at correcting for parameter estimation error; the former based on martingalization techniques and the latter in a non-parametric approach using the generalized cross-spectrum; and Corradi and Swanson (2006a) propose tests robust to both parameter estimation error as well as dynamic misspecification. See Corradi and Swanson (2006b) for an extensive overview of estimation and inference for predictive densities, and Corradi and Swanson (2006c, 2007) for empirical applications.[1]

The main objective of this paper is to provide new methodologies for testing the correct specification of density forecasts that are robust to both dynamic misspecification as well as instabilities. Regarding the robustness to dynamic misspecification, we propose a PIT approach where parameter estimation error is maintained under the null hypothesis, as in Amisano and Giacomini (2007). Maintaining parameter estimation error under the null hypothesis has two advantages: (i) there is no need to correct the statistics for the parameter estimation error, since that is maintained under the null hypothesis; and (ii) the asymptotic distributions of the test statistics for density forecasts at the one-step-ahead horizon are nuisance parameter free even in the presence of dynamic misspecification. Thus, the critical values can be directly tabulated.[2] However, our approach is very different from

[1] See Price (2001) for applications of predictive densities to policymaking.

[2] From a conceptual point of view our tests are concerned only with the distributional assumptions of (potentially misspecified) models since the null hypothesis involves parameter estimates rather than their pseudo-true values. However, the evaluation still involves comparing the density forecast with the true

Amisano and Giacomini (2007): 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. We derive our tests within classes of tests commonly used in the literature, such as the Kolmogorov-Smirnov and Cramér-von Mises-type tests.

Regarding the robustness to instabilities, one of the important assumptions for the validity of the tests proposed by Diebold et al. (1998, 1999), Bai (2003) and Corradi and Swanson (2006a) is stationarity (i.e. absence of structural breaks), which we relax in this paper. In particular, we propose a specification test robust to instabilities by extending the PIT approach to test whether the predictive density is correctly specified *at each point in time*. In addition, we also propose a predictive density instability test that detects distributional change in the predictive densities even if the densities are misspecified.

We show that all our proposed test statistics have good size properties in small samples. In addition, our forecast density test robust to instabilities has good power to detect misspecification of the forecast distribution even when the misspecification affects only a sub-sample.

When misspecification of the predictive density is detected, it is an important step to understand the source of the misspecification, that is whether the null hypothesis is rejected due to the lack of uniformity or independence. Lack of uniformity refers to a situation where, on average, the unconditional probability that the realizations are compatible with the model's predictive density is incorrect. Lack of independence refers to a situation where, even if on average realizations are compatible with the model's predictive density, the pattern of the rejections is non-random. 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. All the tests can be applied to either one-step-ahead or multiple-steps-ahead forecast densities.

Our approach is primarily related to Diebold et al. (1998, 1999) and especially Corradi and Swanson (2006a): we test the null hypothesis of correct specification of density forecasts, although in a way robust to the presence of dynamic misspecification, parameter estimation error, and possibly instability. Our approach is also related to Inoue (2001). Inoue (2001) developed techniques to test whether the in-sample empirical distribution of a model is constant over time. There are two important differences between Inoue's (2001) approach and

realized observations.

ours: we focus on the out-of-sample evaluation of densities (as opposed to in-sample tests) and our null hypothesis is different: it involves testing whether the true predictive distribution matches that implied by a model at each point in time (rather than whether the forecast distribution has changed over time, as in Inoue, 2001). However, we also discuss a modified statistic for testing the constancy of the predictive density over time. Our approach is more distantly related to Rossi (2005): she jointly tests the hypothesis of stability of the parameters as well as that the parameters satisfy a certain restriction in-sample. The approach taken in this paper is similar in that we focus on testing a joint null hypothesis of stability in the predictive distribution as well as correct specification of the predictive distribution. However, it is very different for two reasons: first, because it focuses on prediction, which require a different approach than in-sample tests; second, because it focuses on predictive density tests, which are very different from tests on parameters.

We provide an empirical application of our proposed tests to the density forecasts provided in the Survey of Professional Forecasters (SPF). Our test uncovers that the forecast densities of both output growth and inflation are mis-specified. In addition, we find evidence of instabilities in the forecast densities. Our results indicate that the predictive density of the output growth is mis-specified both before and after the early 1990s, although very differently over time. The implications of the predictive density evaluation for inflation are similar. There are breaks detected for both the density nowcast, as well as the one-year-ahead density forecast of inflation.

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. Section 4 presents results for tests of correct specification of density forecasts robust to both dynamic misspecification and instabilities, and Section 5 discusses how to disentangle lack of uniformity from lack of independence. Section 6 provides Monte Carlo evidence on the performance of our tests in small samples, and Section 7 presents the empirical results. Section 8 concludes.

2 Notation and Definitions

We first introduce the notation and discuss the assumptions about the data, the models and the estimation procedure. We are interested in the true but unknown $h-$step-ahead conditional predictive densities for the scalar variable y_t denoted by $\phi_0(.)$.[3]

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 $h-$steps-ahead out-of-sample density forecasts, such that $R+P-1+h=T+h$. Let $\Im_t$ be the information set at time t and let the sequence of P out-of-sample estimated direct conditional density forecasts be denoted by $\left\{\widehat{\phi}_{t+h}\left(y_{t+h}|\Im_t,\widehat{\theta}_{t,R}\right)\right\}_{t=R}^{T}$, which depend on the in-sample parameter estimates, $\widehat{\theta}_{t,R}$. These parameters are estimated only once, using a sample including data indexed $1,...,R$ (fixed scheme) or re-estimated at each $t=R,...,T$ over a window of R data including data indexed $t-R+1,...,t$ (rolling scheme). In addition to being parametric (such as a normal distribution), the distribution $\widehat{\phi}_{t+h}(.)$ can also be non-parametric (as in the empirical application in this paper).

Consider the probability integral transform (PIT), which is the cumulative density function of $\widehat{\phi}_{t+h}(.)$ evaluated at the realized value y_{t+h}:

$$z_{t+h}=\int_{-\infty}^{y_{t+h}}\widehat{\phi}_{t+h}\left(u|\Im_t,\widehat{\theta}_{t,R}\right)du\equiv\widehat{\Phi}_{t+h}\left(y_{t+h}|\Im_t,\widehat{\theta}_{t,R}\right).$$

Let

$$\xi_{t+h}\equiv\left(1\left\{\widehat{\Phi}_{t+h}\left(y_{t+h}|\Im_t,\widehat{\theta}_{t,R}\right)\leq r\right\}-r\right),$$

and consider $\Psi(r)=\Pr\{z_{t+h}\leq r\}-r$ and its out-of-sample counterpart:

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

where $r\in[0,1]$. Also, let $\pi\in\Pi\subseteq(0,1)$ and

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

In the following sections we first present results for the case of one-step-ahead forecasts, then generalize the tests to the presence of serial correlation. The generalized case could also apply to the $h>1$ step-ahead forecasts. All the proofs are relegated to the Appendix.

[3] 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.

3 Asymptotic Tests of Specification for Predictive Densities in the Presence of Dynamic Misspecification

This section discusses tests for correct specification of a predictive distribution that allow for the presence of dynamic misspecification under the null hypothesis. The tests we propose have an asymptotic distribution that is free of nuisance parameters in the one-step-ahead forecast case, and their critical values can be tabulated. We also discuss tests that are valid in the presence of multi-step-ahead forecasts and serial correlation.

In order to maintain parameter estimation error under the null hypothesis, we state our null hypothesis in terms of estimated parameter values (as in Amisano and Giacomini, 2007). We focus on testing whether $\widehat{\phi}_{t+h}\left(y_{t+h}|\Im_t,\widehat{\theta}_{t,R}\right)=\phi_0\left(y_{t+h}|\Im_t,\widehat{\theta}_{t,R}\right)$, that is:

$$H_0:\widehat{\Phi}_{t+h}\left(y_{t+h}|\Im_t,\widehat{\theta}_{t,R}\right)=\Phi_0\left(y_{t+h}|\Im_t,\widehat{\theta}_{t,R}\right), \tag{3}$$

where $\Phi_0\left(y_{t+h}|\Im_t,\widehat{\theta}_{t,R}\right)\equiv\Pr\left(y_{t+h}\leq y|\Im_t,\widehat{\theta}_{t,R}\right)$. The alternative hypothesis, H_A, is the negation of H_0.

We are interested in the test statistics:

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

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

Note that the κ_P^{CS} 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 $\Psi_P\left(r\right)$, we consider its square). 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: $H_0^{CS}:\widehat{\Phi}_{t+h}\left(y_{t+h}|\Im_t,\theta_0\right)=\Phi_0\left(y_{t+h}|\Im_t,\theta_0\right)$ for some $\theta_0\in\Theta$, where Θ is the parameter space. That is, Corradi and Swanson (2006a) test the hypothesis of correct specification under 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 affects the limiting distribution of their test statistic as well, further contributing to the fact that the limiting distribution depends on nuisance parameters.

Under our null hypothesis (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, and hence parameter estimation error does not affect the asymptotic distribution of the test statistic. Dynamic misspecification is also maintained under the null hypothesis, and by a similar reasoning does not affect the asymptotic distribution of our test statistic.

3.1 One-step-ahead Forecasts

Let $h = 1$. First, we derive the asymptotic distribution of $\Psi_P(r)$ for one-step-ahead forecasts under Assumption 1.

Assumption 1.

(i) $\{y_{t+1}, \Im_t\}_{t=R}^{T}$ is mixing with $\phi(j)$ of size $-\lambda/(2\lambda-1)$ when $\lambda \geq 1$ or $\alpha(j)$ of size $-\lambda/(\lambda-1), \lambda > 1$ and generated from $\left\{\phi_0\left(y_{t+1}|\Im_t, \widehat{\theta}_{t,R}\right)\right\}_{t=R}^{T}$, whose cumulative distribution function $\Phi_0(.)$ is continuous, differentiable and has a well defined inverse;

(ii) $\left\{\widehat{\Phi}_{t+1}\left(y_{t+1}|\Im_t, \widehat{\theta}_{t,R}\right)\right\}_{t=R}^{T}$ has non-zero Jacobian with continuous partial derivatives;

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

Theorem 1 (Asymptotic Distribution of $\Psi_P(r)$) *Under Assumption 1 and H_0 in eq. (3): (i) $\{z_{t+1}\}_{t=R}^{T}$ is i.i.d. $U(0,1)$; (ii) $\Psi_P(r)$ weakly converges (considered as variables in the space $([0,1] \times \mathbb{R})$ to the Gaussian process $\Psi(.)$, with mean zero and 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 distribution of our proposed test statistics have the appealing feature of being nuisance parameter free.

Theorem 2 (Correct Specification Tests) *Under Assumption 1 and H_0 in eq. (8):*

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

and

$$C_P^{CS} \equiv \int \Psi_P(r)' \Psi_P(r)\, dr \Rightarrow \int \Psi(r)' \Psi(r)\, dr. \tag{7}$$

Reject H_0 at the $\alpha \cdot 100\%$ significance level if $\kappa_P^{CS} > \kappa_{\alpha;P}^{CS}$ and $C_P^{CS} > C_{\alpha;P}^{CS}$. Critical values for $\alpha = 10\%$, 5% and 1% are provided in Table 1, Panel A.

INSERT TABLE 1 HERE

3.2 Multi-step-ahead Forecasts

When considering h-steps-ahead forecasts, $h > 1$ and finite, an additional problem arises, as h-steps-ahead forecasts are at least $(h-1)$ serially correlated. Thus, we need to extend our results and allow the forecasts to be both serially correlated and potentially mis-specified under the null hypothesis. Consider the following Assumption:

Assumption 2.

(i) $\{y_{t+h}, \Im_t\}_{t=R}^{T}$ is strong mixing with $\alpha(j)$ of size $-\lambda/(\lambda-1)$,where $\sum_{j=1}^{\infty} j^2 \alpha(j)^{\lambda/(4+\lambda)} < \infty$, $\lambda \in [1,2)$, and generated from $\left\{\phi_0\left(y_{t+h}|\Im_t, \widehat{\theta}_{t,R}\right)\right\}_{t=R}^{T}$, whose cumulative distribution function $\Phi_0(.)$ is continuous, differentiable and has a well defined inverse;

(ii) $\Pr(\xi_{t+h} \leq r_1, \xi_{t+h+d} \leq r_2) = F_d(r_1, r_2)$, where $F_d(.,.)$ and $F(.)$ are the distribution functions of the random variable ξ_{t+h} and $F(.)$ is continuous.

Under serial correlation or h-steps-ahead forecasts, 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.

Theorem 3 (Correct Specification Tests under Serial Correlation) *Under Assumptions 1(ii), 1(iii) and 2 and H_0 in eq. (3), $\Psi_P(r)$ weakly converges (considered as variables in the space $([0,1] \times \mathbb{R})$ to the Gaussian process $\Psi(.)$, with mean zero and auto-covariance function $E\left[\Psi(r_1)' \Psi(r_2)\right] = \sigma(r_1, r_2)$, where $\sigma(r_1, r_2) = \sum_{d=-\infty}^{\infty} [F_d(r_1, r_2) - F(r_1) F(r_2)]$. Furthermore,*

$$\kappa_P^{CS} \Rightarrow \sup_{r \in [0,1]} \left\{\Psi(r)' \Psi(r)\right\},$$

$$C_P^{CS} \Rightarrow \int_0^1 \left\{\Psi(r)' \Psi(r)\right\} dr.$$

For a given estimate of $\sigma(r_1, r_2)$, the critical values of κ_P^{CS} and C_P^{CS} can be obtained via Monte Carlo simulations; alternatively, the asymptotic distributions can be approximated in small samples using a bootstrap procedure.

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 by creating sub-samples of forecast distributions that are at least h-periods apart. However, this procedure may not be possible in small samples, since sub-sampling 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. Alternative approaches include using a block bootstrap to obtain the critical values (Bai and Ng, 2005, and Corradi and Swanson, 2006c). The relative merits and validity of these differential approaches remains to be studied further.

4 Predictive Density Specification Tests in the Presence of Instabilities

This section discusses tests for the correct specification of the forecast density distribution that allow for dynamic misspecification under the null hypothesis and that can detect misspecification in the predictive density even if it arises only in a sub-sample. Again, the tests that we propose have an asymptotic distribution that is free of nuisance parameters in the one-step-ahead forecast case, and their critical values can be tabulated. We also discuss tests that are valid in the presence of multi-step-ahead forecasts and serial correlation.

4.1 One-step-ahead Forecasts

Our interest lies in testing whether $\widehat{\phi}_{t+h}\left(y_{t+h}|\Im_t, \widehat{\theta}_{t,R}\right) = \phi_0\left(y_{t+h}|\Im_t, \widehat{\theta}_{t,R}\right)$ at any point in time t over the out-of-sample portion of the data, that is:

$$H_0 : \widehat{\Phi}_{t+h}\left(y_{t+h}|\Im_t, \widehat{\theta}_{t,R}\right) = \Phi_0\left(y_{t+h}|\Im_t, \widehat{\theta}_{t,R}\right) \text{ for all } t = R, ..., T, \tag{8}$$

where $\Phi_0\left(y_{t+h}|\Im_t, \widehat{\theta}_{t,R}\right) \equiv \Pr\left(y_{t+h} \leq y|\Im_t, \widehat{\theta}_{t,R}\right)$.

The following theorem derives the asymptotic distribution of $\Psi_P(\pi, r)$ for one-step-ahead forecasts under Assumption 1.

Theorem 4 (Asymptotic Distribution of $\Psi_P(\pi, r)$) *Under Assumption 1 and H_0 in eq. (8): (i) $\{z_{t+1}\}_{t=R}^{T}$ is i.i.d. $U(0,1)$; (ii) $\Psi_P(\pi, r)$ weakly converges (considered as variables in the space $\left([0,1]^2 \times \mathbb{R}\right)$ to the Gaussian process $\Psi(.,.)$, with mean zero and auto-covariance function $E\left[\Psi(\pi_1, r_1)\Psi(\pi_2, r_2)\right] = \inf(\pi_1, \pi_2)\left[\inf(r_1, r_2) - r_1 r_2\right]$.*[4]

[4] $\Psi(.,.)$ is a Kiefer process.

Again, note that, since we are evaluating the forecast densities at the estimated parameter values under the null hypothesis, as in Amisano and Giacomini (2007), we do not need to correct the limiting distributions by parameter estimation uncertainty, unlike the approaches in Corradi and Swanson (2006b) and Bai (2003).

We introduce the following notation. Let

$$Q_P(\pi, r) \equiv \left[R_P F_P(\pi, r)\right]' \left[R_P F_P(\pi, r)\right],$$

where

$$F_P(\pi, r) \equiv \left[\begin{array}{c} P^{-1/2} \sum_{t=R}^{R+[\pi P]} \left(1\left\{\widehat{\Phi}_{t+1}\left(y_{t+1}|\Im_t, \widehat{\theta}_{t,R}\right) \leq r\right\} - r\right) \\ P^{-1/2} \sum_{t=R+[\pi P]+1}^{T} \left(1\left\{\widehat{\Phi}_{t+1}\left(y_{t+1}|\Im_t, \widehat{\theta}_{t,R}\right) \leq r\right\} - r\right) \end{array}\right] \tag{9}$$

$$R_P = \left[\begin{array}{cc} (1-\pi) & -\pi \\ 1 & 1 \end{array}\right] \tag{10}$$

Again, we consider two types of test statistics: the first is a weighted Kolmogorov-Smirnov-type statistic and the second is a weighted Cramér-von Mises-type statistic:

$$\kappa_P \equiv \sup_{\pi \in \Pi} \sup_{r \in [0,1]} Q_T(\pi, r) \tag{11}$$

$$C_P \equiv \textstyle\int_\pi \int_r Q_T(\pi, r)\, d\pi dr. \tag{12}$$

Theorem 5 (Forecast Density Tests Robust to Instabilities) *Under Assumption 1 and H_0 in eq. (8):*

$$\begin{aligned} \kappa_P &= \sup_{\pi \in \Pi} \sup_{r \in [0,1]} Q_P(\pi, r) \\ &\Rightarrow \sup_{\pi \in \Pi} \sup_{r \in [0,1]} \left\{\Psi^\circ(\pi, r)' \Psi^\circ(\pi, r) + \Psi(1, r)' \Psi(1, r)\right\}, \end{aligned}$$

and

$$C_P \equiv \textstyle\int_\Pi \int_0^1 Q_P(\pi, r)\, d\pi dr \Rightarrow \int_\Pi \int_0^1 \left\{\Psi^\circ(\pi, r)' \Psi^\circ(\pi, r) + \Psi(1, r)' \Psi(1, r)\right\} d\pi dr.$$

where $\Psi^\circ(\pi, r) \equiv \Psi(\pi, r) - \pi \Psi(1, r)$ is a Gaussian process with zero mean and covariance function $E\{\Psi^\circ(\pi_1, r_1)\Psi^\circ(\pi_2, r_2)\} = [\inf(\pi_1, \pi_2) - \pi_1\pi_2][\inf(r_1, r_2) - r_1 r_2]$ (see Deshayes and Picard (1986), p.127). Reject H_0 at the $\alpha \cdot 100\%$ significance level if $\kappa_P > \kappa_{\alpha;P}$ and $C_P > C_{\alpha;P}$. Critical values for $\alpha = 10\%$, 5% and 1% are provided in Table 1, Panel A for $\Pi = [0.15, ..., 0.85]$.

Note that one could be interested in testing for correct specification in specific parts of the distribution.[5] For example, one might be interested in the tails of the distribution, which correspond to outliers, for example the left tail where $r \in [0, 0.25]$, or the right tail where $r \in [0.75, 1]$, or both: $r \in \{[0, 0.25] \cup [0.75, 1]\}$. Or, 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 Panel B of Table 1.

Note that our test is, by construction, robust to dynamic misspecification. We demonstrate this property in our Monte Carlo simulations as well.

Our proposed tests differ substantially from those existing in the literature. In particular, the test proposed by Corradi and Swanson (2006a) would be a special case of our approach for $R_P = \left[\begin{array}{cc} 1 & 1 \end{array}\right]$ instead of eq. (10). Thus, there are two differences between our paper and theirs: first of all the null hypothesis is different. Second, we are concerned with testing for the correct specification of the predictive density in the presence of instabilities whereas they assume stationarity.[6]

Note that our tests are also different from Inoue (2001), who tests the null hypothesis of constancy of in-sample densities over time. His null hypothesis in our out-of-sample context is:

$$H_0^I : [\pi P]^{-1/2} \sum_{t=1}^{[\pi P]} I\left(z_{t+h} \leq r\right) - \left(P - [\pi P]\right)^{-1/2} \sum_{t=[\pi P]+1}^{P} I\left(z_{t+h} \leq r\right) = 0,$$

for $\pi \in \Pi \subseteq (0,1)$. Thus, Inoue (2001) is another special case of our approach where $R_P = \left[\begin{array}{cc} (1-\pi), & -\pi \end{array}\right]$ instead of eq. (10). In other words, our proposed κ_P and C_P statistics focus on testing the *joint* null hypothesis of constancy of forecast densities over time as well as their correct specification. Again, because our null hypothesis is specified in terms of the estimated parameters, parameter estimation error does not affect the asymptotic distribution of our tests. In contrast, in Inoue (2001) parameter estimation error matters because he focuses on an in-sample test where the null hypothesis is evaluated at the true parameter values.

At the same time, one could consider tests for instabilities in the forecast densities inspired by Inoue (2001): the Kolmogorov-Smirnov test is:

$$\kappa_P^I = \sup_{\pi \in \Pi} \sup_{r \in [0,1]} \left\{ \left[[\pi P]^{-1/2} \sum_{t=1}^{[\pi P]} I\left(z_{t+h} \leq r\right) - \left(P - [\pi P]\right)^{-1/2} \sum_{t=[\pi P]+1}^{P} I\left(z_{t+h} \leq r\right) \right] \left(\frac{\tau}{P}\right)\left(1 - \frac{\tau}{P}\right) P^{-1/2} \right\}^2,$$

[5] 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.

[6] In other words, we test the null hypothesis that the PIT is uniform at each point in time, and we maintain parameter estimation error under the null hypothesis.

and the Cramér-von Mises test statistic is instead:

$$C_P^I = \int_{\Pi} \int_0^1 \left\{ \left[[\pi P]^{-1/2} \sum_{t=1}^{\tau} I\left(z_{t+h} \leq r\right) - \left(P - [\pi P]\right)^{-1/2} \sum_{t=[\pi P]+1}^{P} I\left(z_{t+h} \leq r\right) \right] \pi \left(1-\pi\right) P^{-1/2} \right\}^2 d\pi dr.$$

We provide critical values for the statistics κ_P^I and C_P^I for testing the constancy of the forecast density over time for $h = 1$ in Table 1.

4.2 Multi-step-ahead Forecasts

As previously discussed, in the case of multi-step-ahead forecasts are serially correlated by construction. Theorem 6 provides the test statistics that can be used in such situations.

Theorem 6 (Density Tests Robust to Instabilities under Correlation) *Under Assumptions 1(ii), 1(iii) and 2 and H_0 in eq. (8):*
(a) $\Psi_P(\pi, r)$ weakly converges (considered as variables in the space $\left([0,1]^2 \times \mathbb{R}\right)$ to the Gaussian process $\Psi(.,.)$, with mean zero and auto-covariance function $E\left[\Psi(\pi_1, r_1)' \Psi(\pi_2, r_2)\right] = \inf(\pi_1, \pi_2)\, \sigma(r_1, r_2)$, where $\sigma(r_1, r_2) = \sum_{d=-\infty}^{\infty} \left[F_d(r_1, r_2) - F(r_1) F(r_2)\right]$;
(b)

$$\begin{aligned} \kappa_P &= \sup_{\pi \in \Pi} \sup_r Q_P(\pi, r) \qquad (13) \\ &\Rightarrow \sup_{\pi \in \Pi} \sup_r \left\{ \Psi^{\circ}(\pi, r)' \Psi^{\circ}(\pi, r) + \widetilde{\Psi}(1, r)' \Psi(1, r) \right\}, \end{aligned}$$

$$C_P \equiv \int_{\pi}\int_{r} Q_P(\pi, r)\, d\pi dr \Rightarrow \int_{\pi}\int_{r} \left\{ \Psi^{\circ}(\pi, r)' \Psi^{\circ}(\pi, r) + \Psi(1, r)' \Psi(1, r) \right\} d\pi dr, \qquad (14)$$

where $\Psi^{\circ}(\pi, r) \equiv \Psi(\pi, r) - \pi \Psi(1, r)$.

For a given estimate of $\sigma(r_1, r_2)$, the critical values of $\Psi(.,.)$ can be obtained via Monte Carlo simulations; alternatively, the asymptotic distribution can be approximated in small samples using a bootstrap procedure similar to that described in Inoue (2001).

5 Understanding the Sources of the Misspecification

When our test rejects, the rejection could be indicating either a violation of independence, uniformity or identical distribution. To evaluate the sources of the rejection, we consider the following tests.

First, we suggest to use a test for uniformity robust to violations of independence, which detects violations of uniformity even if the forecasts were serially correlated. If the researcher is not worried about instabilities in the sample, (s)he could use the test statistics described in Theorem 3. If the researcher is worried about instabilities, (s)he could test for uniformity in a way robust to violation of independence even if it happens in a sub-sample by using the test statistics described in Theorem 5.

Second, 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 tests detect auto-correlation in a linear framework whereas the BDS test is a non-parametric test of independence and identical distribution against an unspecified alternative.

6 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 mis-specified forecasting models.

6.1 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 of $R = [50, 10, 200]$ and $P = [50, 100, 200, 1000]$ to evaluate the performance of the proposed procedure in finite samples. The DGPs are the following:

DGP S1 (Baseline Model): Let $y_t = \mu + \varepsilon_t$, where $\varepsilon_t \sim iidN(0, 1)$ and $\mu = 5$.[7] We estimate a model with a constant, $\widehat{\mu}_t = R^{-1} \sum_{j=t-R+1}^{t} y_j$, $t = R, ..., T$ and use it to produce one-step-ahead forecasts. To ensure that the null hypothesis (8) holds, namely that the predictive density evaluated at the estimated (as opposed to true) parameter values is Uniform, we generate the data under the null hypothesis according to $\widetilde{y}_{t+1} = \widehat{\mu}_t + \eta_{t+1}$, where $\eta_t \sim iidN(0, 1)$ independent of ε_t.

DGP S2 (Estimated Model): Let $y_t = \mu + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \varepsilon_t$. We parameterize the model according to the realistic situation where the researcher is interested

[7] The results are unchanged if a different value of μ is considered.

in forecasting one-quarter-ahead real GDP growth (y_t) with a bivariate model that includes the term spread (x_t) in U.S. data from 1959:I-2010:III. The lag length is selected by BIC in-sample: we first choose the best-fitting autoregressive model by BIC, then augment it with the optimal lags of the variable x_t selected again by BIC. The unconditional means of output growth and term spread are estimated from the data, and are 3.09 and 0.92 respectively. The process for the term spread (x_t) is an AR(p) where the number of lags equals 1. That is, $x_t = 0.2 + 0.8x_{t-1} + \nu_t$, where $\nu_t \sim iidN(0, 1.08^2)$. The remaining parameters are estimated as: $\mu = 1.50, \beta_1 = 0.20, \beta_2 = -0.22, \gamma_1 = 0.13, \gamma_2 = 0.82$. In addition, $\varepsilon_t \sim iidN\,(0, 9.54^2)$. The forecasts are generated under the null hypothesis and are constructed under correct specification: $\widetilde{y}_{t+1} = \hat{\mu} + \widehat{\beta}_{1,t} y_t + \widehat{\beta}_{2,t} y_{t-1} + \widehat{\gamma}_{1,t} x_t + \widehat{\gamma}_{2,t} x_{t-1} + \eta_{t+1}$, where the parameters are estimated by OLS in rolling windows.

The estimated model in both DGPs S1 and S2 are correctly specified. We consider also a DGP where the model is mis-specified:

DGP S3 (Misspecification): The data generating process is exactly that of DGP S2, except the null hypothesis is constructed under misspecification: $\widetilde{y}_{t+1} = \hat{\mu} + \hat{\beta}_{2,t} y_{t-1} + \hat{\gamma}_{2,t} x_{t-1} + \eta_{t+1}$.

DGPs S1-S3 are based on one-step-ahead forecast densities. To investigate the case of h-steps ahead forecast densities and the presence of serial correlation, we consider DGPS4.

DGP S4 (Serial Correlation): The data is generated by $y_t = \mu + \varepsilon_t + \rho\varepsilon_{t-1}$, where $\rho = 0.2$ and $\varepsilon_t \sim iidN\,(0, 1)$. The null hypothesis is constructed based on $\widetilde{y}_t = \hat{\mu} + \varepsilon_t + \rho\varepsilon_{t-1}$, where $\varepsilon_t \sim iidN\,(0, \widehat{\sigma}^2)$.

The results are shown in Tables 2 and 3. Panel A in Table 2 considers the forecast density test robust to instabilities (κ_P, C_P); panel B considers tests for the correct specification of the forecast density (κ_P^{CS}, C_P^{CS}); and panel C considers tests for instabilities in the forecast densities over time (κ_P^{I}, C_P^{I}). Table 2 shows that all our test performs very well in finite samples, with only very mild under-rejections for small values of R and P for the Kolmogorov-Smirnov-type test. Table 3 shows that in the case of serial correlation, the asymptotic distribution of the Kolmogorov-Smirnov-type test in Theorem 6 approximated using HAC-consistent variance estimates tends to over-reject in finite samples, although mildly; the Cramér-von Mises-type test instead performs fairly well.

INSERT TABLES 2 AND 3 HERE

6.2 Power Analysis

To investigate the power properties of our tests, we consider three DGPs. In each of the DGPs, the researcher evaluates whether the forecasts are compatible with a normal distribution. The DGPs are:

DGP P1: The data are generated from a χ_1^2 distribution, and the researcher tests whether it is normal: $\widetilde{y}_t = \widehat{\mu}_t + (1-c)\,\eta_{1,t} + c\left(\eta_{2,t}^2 - 1\right)\sqrt{2}$, where $\eta_{1,t}$ and $\eta_{2,t}$ are $iidN\,(0,1)$ and independent of each other;[8]

DGP P2: The data are generated from a normal distribution whose parameters change over time: $y_t = \mu_t + \varepsilon_t$, where $\varepsilon_t \sim iidN\,(0, \sigma_t)$ and $\mu_t = \mu_1 = 5$, $\sigma_t = \sigma_1 = 1$ for $t = 1, 2, ..., T - T_1$ and $\mu_t = c\mu_1$, $\sigma_t = c\sigma_1$ for $t = T - T_1 + 1, ..., T$; T_1 is chosen to be at the middle of the out of sample period P. We report results for various values of $c = 0, 0.01, ..., 0.03$.

DGP P3: The shape of the distribution in the data changes over time from a chi-square to a normal: $\widetilde{y}_t = \widehat{\mu}_t + \eta_{1,t} \cdot 1\,(t > T_1) + \left[\left(\eta_{2,t}^2 - 1\right)\sqrt{2}\right] \cdot 1\,(t \leq T_1)$, where $\eta_{1,t}$ and $\eta_{2,t}$ are $iidN\,(0,1)$, independent of each other.

The results are shown in Table 4. The table shows that, across all designs, our proposed density forecast tests robust to instabilities (κ_P, C_P) have good power properties in detecting misspecification in the forecast density, even when it only appears in parts of the sample. On the other hand, the other tests may lack power: Panel A shows that the instability tests (κ_P^I, C_P^I) have no power to detect misspecification in the forecast density when the misspecification is constant over time; the correct specification tests $(\kappa_P^{CS}, C_P^{CS})$ do instead detect misspecification, and have higher power than the density forecast tests robust to instabilities. At the same time, Panel B shows that the correct specification tests have no power to detect mis-specifications in the forecast distribution in the presence of instabilities, although the instability tests would detect instabilities, and with a higher power than the density tests robust to instabilities.

INSERT TABLE 4 HERE

[8]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.

7 Evaluation of the SPF Density Forecasts

Diebold et al. (1999) evaluate density forecasts of inflation based on the mean probability forecasts provided in the Survey of Professional Forecasters (SPF). In particular, they assess whether realized inflation rates are consistent with the description of uncertainty implied by the mean probability distribution forecasts of the survey. Interestingly, the authors note the presence of time variation and emphasize that in their sample the distribution has shifted from overestimating a large negative shock to overestimating large shocks of either sign. They provide results based on sub-sample analysis where the break date is chosen exogenously.

We conduct a formal test of correct specification for the SPF forecast densities taking into account the presence of time variation. In addition to inflation, we also investigate the conditional density forecasts of output growth. We use real GNP/GDP and GNP/GDP deflator as measures of output and prices. The mean probability distribution forecasts are obtained from the Survey of Professional Forecasters, which are publicly available from the Federal Reserve Bank of Philadelphia. The realized values of inflation and output growth are based on the real-time data set for macroeconomists, again, available through the Federal Reserve Bank of Philadelphia.[9] In this 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. We base our results on a continuous approximation with a fitted normal distribution.

The data on the probability distribution forecasts is complicated since the 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 eliminate these problematic issues, we truncate the data set and consider the mean probability distribution forecasts for the period 1981:III-2009:IV.[10] We use the year-over-year growth rates from the second data revision to evaluate the density forecasts. In order to do so, we only look at the first quarter vintage of the real GNP/GDP and GNP/GDP deflator, and calculate the year-over-year growth

[9] The data are available at http://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/ and http://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/.

[10] The changes in the interval range for the real GDP in 2009 have been re-adjusted to make it consistent with that of the year before.

rates for the past year. For example, the growth rate of real output for 1981 is obtained by looking at the 1982:I vintage of the data and applying the following transformation: $100[\ln(y_{1981:IV}) - \ln(y_{1980:IV})]$, where $y_{1981:IV}$ denotes the value of the GDP in the fourth quarter of 1981. We test whether the density nowcasts and one-step-ahead forecasts of output growth and inflation provided in the SPF are in fact normal with a mean and variance obtained with a normal kernel fit over the full sample.

The empirical results are shown in Table 5. The table shows three different results. Panel A considers the results of the tests of correct specification robust to instabilities proposed in this paper (the κ_P, C_P tests). Panel B considers the null hypothesis of correct forecast specification only (the κ_P^{CS}, C_P^{CS} tests), while Panel C considers the hypothesis of instabilities only (the κ_P^I, C_P^I tests). The test statistics consistently reject the respective null hypotheses under both independence and serial correlation. The only exception is the nowcast of inflation. In this case the instability-only test fails to reject the null hypothesis of time variation under independence; however, it rejects it when we allow for serial correlation under the null. In other words, once we allow for serial correlation under the null, we detect changes in the distribution of the PITs over time with all our test statistics, and we reject the correct specification of the forecast density in both output growth and inflation.

INSERT TABLE 5 HERE

Our results are important in the light of the finding that survey forecasts are reportedly providing the best forecasts of inflation. For example, Ang, Bekaert and Wei (2007) find that survey forecsts 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 are not providing a correct forecast for the whole distribution of inflation.

Panel A in Figure 1 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 crude approximation with a binomial distribution as in Diebold, Tay and Wallis (1999).[11] The figure gives a visual representation for the misspecification in the PITs of the output growth: both density nowcast and one-year-ahead density forecasts are mis-specified, and suggest that the survey typically underpredicts future large realizations of output growth.

[11] Since the binomial is a discrete distribution, the probability values at the bands might deviate from being exactly 0.025 and 0.975. The bands indicate statistical significance for each bin separately, not jointly.

Figure 2, Panel A, shows instead the PITs for inflation. The density nowcast, depicted on the left, seems to be mis-specified: it overpredicts large positive and negative surprises in inflation. However the one-year-ahead density forecast, depicted on the right, seems to be closer to a uniform distribution with mild underprediction of a large positive surprise.

INSERT FIGURES 1 AND 2 HERE

The joint test as well as the instability tests indicate a potential break. Though in this paper we are not concerned with the consistency or the uncertainty associated with the break date, we still take the break dates implied by the joint test and consider visual plots of the PITs over the sub-samples to get further insights. The joint test suggests a break in 1991:IV and 1990:IV for the density nowcasts and one-year-ahead forecasts of output growth, respectively, which are roughly consistent with the date the survey switched from collecting GNP figures to GDP figures. On the other hand, when predicting inflation, the implied breaks are around 1985:III for the nowcast, and 1998:II for the one-year-ahead inflation forecast. We then divide the sample according to the likely time of the break, and then plot the PITs in the corresponding sub-samples.

Panel B in Figure 1 plots the PITs for output growth before and after the conjectured break date. Interestingly, Panel B shows that, in the case of output growth nowcasts, the rejections affect both sub-samples, although very differently: before 1992, the rejection is due to the fact that too many realizations fall in the middle of the sample than those expected based on normality. After 1992, forecasters seem to underpredict output growth, as too many realizations end up being in the tail than an estimated normal distribution would have suggested. Looking at the one-year-ahead forecast of output growth, depicted in Panel C, the misspecification appears to affect the forecasts before and after the break in a roughly similar way: there are many realization in the middle of the distribution and in the right tail. The behavior of inflation is different. The PITs for the inflation density nowcast appear to have moved from having many surprise realizations in the middle of the distribution to being better calibrated after 1985:III. The PITs for one-year-ahead forecasts, on the other hand, show an increasing degree of misspecification after 1998, leading to heavy underprediction of low levels of inflation.

8 Conclusions

This paper proposes new tests for predictive density evaluation. They are designed to maintain parameter estimation error under the null hypothesis and are robust to the presence of mis-specification as well as instabilities. The techniques are based on Kolmogorov-Smirnov and Cramér-von Mises-type test statistics. We provide critical values for the test statistics on the whole density forecasts as well as test statistics that focus on specific parts of the density. An empirical application of the proposed methodologies to the Survey of Professional Forecasters uncovers that both their output growth and inflation density forecast are misspecified and finds significant evidence of time-variation in the density forecasts.

References

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

[2] 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.

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

[4] 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.

[5] 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.

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

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

[8] 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.

[9] Corradi, V. and N. R. Swanson (2006c), "Predictive density and conditional confidence interval accuracy tests", *Journal of Econometrics* 135(1–2), 187-228.

[10] Valentina Corradi, Norman R. Swanson (2007), "Evaluation of Dynamic Stochastic General Equilibrium Models Based on Distributional Comparison of Simulated and Historical Data", *Journal of Econometrics* 136(2), 699-723.

[11] Deshayes, J. and D. Picard (1986), "Off-line Statistical Analysis of Change Point Models Using Non Parametric and Likelihood Method". In M. Basseville and A. Benveniste (eds.), *Detection of Abrupt Changes in Signals and Dynamical Systems*, Lecture Notes in Control and Information Sciences 77, 103-168, Berlin: Springer-Verlag.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Hong, Y.M. and H. Li (2003), "Out-of-sample Performance of Spot Interest Rate Models", *mimeo.*

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

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

[19] Newey, W. and K. West (1987), “A Simple, Positive Semi-Definite, Heteroskedasticity and Auto-correlation Consistent Covariance Matrix”, *Econometrica* 55, 703-708.

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

[21] Price, S. (2011), “Forecast Densities and Communicating Uncertainty”, *Bank of England Research Newsletter*.

[22] Rosenblatt, M. (1952), “Remarks on a Multivariate Transformation”, *Annals of Mathematical Statistics* 23, 470-472.

[23] Rossi, B. (2005), “Optimal Tests for Nested Model Selection with Underlying Parameter Instabilities”, *Econometric Theory* 21(5), 962-990.

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

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

9 Appendix A. Proofs

First, we will prove Theorems 4, 5 and 6. The proof of Theorems 1, 2 and 3 follow as a special cases.

Proof of Theorem 4. (i) The joint predictive density of $\{y_{t+1}\}_{t=R}^{T}$ can be decomposed as: $\widehat{\phi}\left(y_{T+1}, ..., y_{R+1}|\Im_R\right) = \widehat{\phi}_{T+1}\left(y_{T+1}|\Im_T\right)\widehat{\phi}_T\left(y_T|\Im_{T-1}\right)...\widehat{\phi}_{R+1}\left(y_{R+1}|\Im_R\right)$. Let $q\left(z_{R+1}, ..., z_{T+1}\right)$ denote the joint density of the probability integral transforms. By using the change of variables formula,

$$\begin{aligned}
q\left(z_{R+1}, ..., z_{T+1}\right) &= \left|\begin{array}{ccc} \left(\partial y_{R+1}/\partial z_{R+1}\right) & ... & \left(\partial y_{R+1}/\partial z_{T+1}\right) \\ ... & ... & ... \\ \left(\partial y_{T+1}/\partial z_{R+1}\right) & ... & \left(\partial y_{T+1}/\partial z_{T+1}\right) \end{array}\right| \times \\
&\quad \times \phi_0\left(\widehat{\Phi}_{T+1}^{-1}\left(z_{T+1}\right)|\Im_T\right)\phi_0\left(\widehat{\Phi}_T^{-1}\left(z_T\right)|\Im_{T-1}\right)...\phi_0\left(\widehat{\Phi}_{R+1}^{-1}\left(z_{R+1}\right)|\Im_R\right) \\
&= \left(\partial y_{R+1}/\partial z_{R+1}\right)...\left(\partial y_T/\partial z_T\right)\left(\partial y_{T+1}/\partial z_{T+1}\right) \times \\
&\quad \times \phi_0\left(\widehat{\Phi}_{T+1}^{-1}\left(z_{T+1}\right)|\Im_T\right)\phi_0\left(\widehat{\Phi}_T^{-1}\left(z_T\right)|\Im_{T-1}\right)...\phi_0\left(\widehat{\Phi}_{R+1}^{-1}\left(z_{R+1}\right)|\Im_R\right),
\end{aligned}$$

where the last equality holds because the Jacobian is lower triangular provided we are in a conditional forecasting framework.

Then since $\left(\frac{1}{\partial z_{R+1}/\partial y_{R+1}}\right) = \frac{1}{\widehat{\phi}_{R+1}\left(\widehat{\Phi}_{R+1}^{-1}\left(z_{R+1}\right)\right)}$,

$$\begin{aligned}
q\left(z_{R+1}, ..., z_{T+1}\right) &= \frac{\phi_0\left(\widehat{\Phi}_{R+1}^{-1}\left(z_{R+1}\right)|\Im_R\right)}{\widehat{\phi}_{R+1}\left(\widehat{\Phi}_{R+1}^{-1}\left(z_{R+1}\right)\right)} \times ... \times \frac{\phi_0\left(\widehat{\Phi}_T^{-1}\left(z_T\right)|\Im_{T-1}\right)}{\widehat{\phi}_T\left(\widehat{\Phi}_T^{-1}\left(z_T\right)\right)} \times \\
&\quad \times \frac{\phi_0\left(\widehat{\Phi}_{T+1}^{-1}\left(z_{T+1}\right)|\Im_T\right)}{\widehat{\phi}_{T+1}\left(\widehat{\Phi}_{T+1}^{-1}\left(z_{T+1}\right)\right)}.
\end{aligned}$$

Under $H_0 : \widehat{\phi}_{t+1}\left(y_{t+1}|\Im_t, \widehat{\theta}_{t,R}\right) = \phi_0\left(y_{t+1}|\Im_t, \widehat{\theta}_{t,R}\right)$ then z_t is $U\left(0,1\right)$ (see Theorem 2.1.10 in Casella and Berger, 2002) and given that each of the ratios are the probability distribution of the corresponding z (since, as shown by Diebold et al. (1998), the distribution of z_{R+1}, by the change of variable formula, is $\left|\frac{\partial\left[\Phi_0\left(z_{R+1}\right)\right]^{-1}}{\partial z_{R+1}}\right|\widehat{\phi}\left(\left[\Phi_0\left(z_{R+1}\right)\right]^{-1}\right) = \frac{\phi_0\left(\left[\Phi_0\left(z_{R+1}\right)\right]^{-1}\right)}{\widehat{\phi}\left(\left[\Phi_0\left(z_{R+1}\right)\right]^{-1}\right)}$) then their product is a multivariate $U\left(0,1\right)$ distribution for $\{z_{t+1}\}_{t=R}^{T}$, which implies that $\{z_{t+1}\}_{t=R}^{T}$ is $iidU\left(0,1\right)$.

(ii) Under H_0, z_{t+1} is uniformly distributed on $[0,1]$. Then, from Proposition 1 in Shorack and

Wellner (1986, p. 131), $\Psi_P(\pi, r) = P^{-1/2} \sum_{t=R}^{R+[\pi P]} \left(1\left\{\widehat{\Phi}_{t+1}\left(y_{t+1}|\Im_t, \widehat{\theta}_{t,R}\right) \le r\right\} - r\right)$ weakly converges (considered as variables in the space $\left([0,1]^2 \times \mathbb{R}\right)$ to the Kiefer process $\Psi^\circ(.,.)$, with mean zero and auto-covariance function $E\left[\Psi^\circ(\pi_1, r_1)\Psi^\circ(\pi_2, r_2)\right] = \inf(\pi_1, \pi_2)\left[\inf(r_1, r_2) - r_1 r_2\right]$.
■

Proof of Theorem 5. Eqs. (9) and (10) imply:

$$
\begin{aligned}
R_P F_P(\pi, r) &\equiv \begin{bmatrix} \left[(1-\pi) P^{-1/2} \sum_{t=R}^{R+[\pi P]} \xi_{t+1} - \pi P^{-1/2} \sum_{t=R+[\pi P]+1}^{T} \xi_{t+1}\right] \\ P^{-1/2}\left[\sum_{t=R}^{R+[\pi P]} \xi_{t+1} + \sum_{t=R+[\pi P]+1}^{T} \xi_{t+1}\right] \end{bmatrix} \\
&= \begin{bmatrix} \left[P^{-1/2} \sum_{t=R}^{R+[\pi P]} \xi_{t+1} - P^{-1/2} \sum_{t=R+[\pi P]+1}^{T} \xi_{t+1}\right] \\ P^{-1/2} \sum_{t=R}^{T} \xi_{t+1} \end{bmatrix} \\
&\Rightarrow \begin{pmatrix} \left[\Psi(\pi, r) - \pi\Psi(1, r)\right] \\ \Psi(1, r) \end{pmatrix} = \begin{pmatrix} \Psi^\circ(\pi, r) \\ \Psi(1, r) \end{pmatrix}.
\end{aligned}
$$

Thus, by the Continuous Mapping theorem,

$$
\begin{aligned}
&\left[R_P F_P(\pi, r)\right]' \left[R_P F_P(\pi, r)\right] \\
&\Rightarrow \Psi^\circ(\pi, r)' \Psi^\circ(\pi, r) + \Psi(1, r)' \Psi(1, r).
\end{aligned}
$$

■

Proof of Theorem 6. Theorem 6 follows from Inoue (2001), since $\Psi_P(\pi, r)$ weakly converges (considered as variables in the space $\left([0,1]^2 \times \mathbb{R}\right)$) to Gaussian process $\Psi(.,.)$. ■

Proof of Theorem 1. The proof of Theorem 1 follows from the proof of Theorem 4 by setting $\pi_1 = \pi_2 = 1$. ■

Proof of Theorem 2. The proof of Theorem 2 follows from the proof of Theorem 5 by setting $\pi_1 = \pi_2 = 1$. ■

Proof of Theorem 3. Theorem 3 is a special case of Theorem 6 when $\pi_1 = \pi_2 = 1$. ■

10 Appendix B. Tables and Figures

Table 1. Critical Values

	α :	0.01	0.05	0.10	0.01	0.05	0.10
Panel A. Tests on the Whole Distribution							
Density Test Robust to Instabilities ($\kappa_{\alpha;P}$,$C_{\alpha;P}$)		2.4484	1.6720	1.3699	0.7648	0.4753	0.3759
Correct Specification Test ($\kappa^{CS}_{\alpha;P}$, $C^{CS}_{\alpha;P}$)		2.2685	1.4792	1.1552	0.7286	0.4440	0.3353
Instability Test ($\kappa^{I}_{\alpha;P}$, $C^{I}_{\alpha;P}$)		0.7785	0.5860	0.5050	0.1193	0.0773	0.0617
Panel B. Tests on Specific Parts of the Distribution							
Density Tests Robust to Instabilities			$\kappa_{\alpha;P}$			$C_{\alpha;P}$	
Right Tail	$r \in (0, 0.25]$	1.3078	0.8662	0.6523	0.5481	0.3237	0.2367
Right Half	$r \in (0, 0.50]$	2.2168	1.5447	1.2243	0.8793	0.5719	0.4176
Left Half	$r \in [0.50, 1)$	2.3026	1.5248	1.2348	0.9710	0.5814	0.4463
Left Tail	$r \in [0.75, 1)$	1.5355	0.9984	0.7666	0.7494	0.4282	0.3219
Center	$r \in [0.25, 0.75]$	2.3751	1.7033	1.3818	1.1399	0.7412	0.5453
Tails	$r \in \{(0, 0.25] \cup [0.75, 1)\}$	1.6140	1.1267	0.9100	0.4711	0.3120	0.2504
Correct Specification Tests			$\kappa^{CS}_{\alpha;P}$			$C^{CS}_{\alpha;P}$	
Right Tail	$r \in (0, 0.25]$	1.1552	0.8200	0.5202	0.5202	0.3038	0.2163
Right Half	$r \in (0, 0.50]$	2.0000	1.3448	1.0082	0.8368	0.5401	0.3802
Left Half	$r \in [0.50, 1)$	2.0808	1.3122	1.0368	0.9387	0.5409	0.4017
Left Tail	$r \in [0.75, 1)$	1.3448	0.8450	0.6050	0.7252	0.4031	0.2935
Center	$r \in [0.25, 0.75]$	2.1632	1.5138	1.1552	1.1029	0.6877	0.4983
Tails	$r \in \{(0, 0.25] \cup [0.75, 1)\}$	1.4621	0.9800	0.7688	0.4466	0.2877	0.2275
Instability Tests			$\kappa^{I}_{\alpha;P}$			$C^{I}_{\alpha;P}$	
Right Tail	$r \in (0, 0.25]$	0.4484	0.3080	0.2517	0.0781	0.0494	0.0372
Right Half	$r \in (0, 0.50]$	0.7021	0.5139	0.4380	0.1240	0.0806	0.0630
Left Half	$r \in [0.50, 1)$	0.7360	0.5367	0.4447	0.1364	0.0890	0.0691
Left Tail	$r \in [0.75, 1)$	0.4956	0.3446	0.2889	0.1044	0.0641	0.0484
Center	$r \in [0.25, 0.75]$	0.7696	0.5918	0.4978	0.1770	0.1122	0.0875
Tails	$r \in \{(0, 0.25] \cup [0.75, 1)\}$	0.5213	0.3853	0.3284	0.0687	0.0478	0.0390

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 domains for π and r are discretized with a step sizes of 0.002 and 0.05, respectively.

Table 2. Size Properties - I.I.D. Case

Panel A: Forecast Density Test Robust to Instabilities

DGP S1

		κ_P			C_P		
P	$R:$	50	100	200	50	100	200
50		0.05	0.05	0.05	0.06	0.06	0.06
100		0.05	0.06	0.05	0.06	0.06	0.05
200		0.05	0.05	0.05	0.06	0.05	0.06
500		0.05	0.05	0.06	0.06	0.06	0.06
1000		0.05	0.06	0.05	0.06	0.06	0.06

DGP S2

		κ_P			C_P		
P	$R:$	50	100	200	50	100	200
50		0.05	0.05	0.05	0.06	0.06	0.06
100		0.05	0.05	0.05	0.06	0.06	0.05
200		0.05	0.05	0.05	0.06	0.05	0.06
500		0.05	0.06	0.05	0.05	0.06	0.06
1000		0.06	0.06	0.05	0.06	0.06	0.06

DGP S3

		κ_P			C_P		
P	$R:$	50	100	200	50	100	200
50		0.05	0.05	0.05	0.06	0.05	0.06
100		0.05	0.06	0.05	0.06	0.06	0.06
200		0.06	0.06	0.05	0.06	0.06	0.05
500		0.06	0.05	0.06	0.06	0.05	0.06
1000		0.05	0.06	0.06	0.06	0.06	0.06

Note. The table reports empirical rejection frequencies for the test statistics κ_P and C_P at the 5% nominal size for various values of P and R. The number of Monte Carlo replications is 5,000. The domains for π and r are discretized with a step sizes of 0.002 and 0.05, respectively.

Panel B: Correct Specification Tests

DGP S1							
		κ_P^{CS}			C_P^{CS}		
P	$R:$	50	100	200	50	100	200
50		0.05	0.04	0.04	0.06	0.06	0.06
100		0.04	0.05	0.04	0.06	0.06	0.05
200		0.05	0.05	0.05	0.06	0.05	0.06
500		0.05	0.05	0.05	0.05	0.06	0.06
1000		0.05	0.06	0.05	0.06	0.06	0.05

DGP S2							
		κ_P^{CS}			C_P^{CS}		
P	$R:$	50	100	200	50	100	200
50		0.05	0.05	0.05	0.06	0.05	0.06
100		0.04	0.04	0.04	0.06	0.05	0.05
200		0.04	0.05	0.05	0.06	0.05	0.06
500		0.05	0.05	0.05	0.05	0.06	0.06
1000		0.05	0.06	0.05	0.06	0.06	0.05

DGP S3							
		κ_P^{CS}			C_P^{CS}		
P	$R:$	50	100	200	50	100	200
50		0.04	0.04	0.05	0.06	0.05	0.06
100		0.05	0.05	0.04	0.06	0.06	0.05
200		0.05	0.05	0.04	0.06	0.06	0.05
500		0.05	0.05	0.05	0.06	0.05	0.06
1000		0.06	0.06	0.05	0.06	0.06	0.06

Note. The table reports empirical rejection frequencies for the test statistics κ_P^{CS} and C_P^{CS} 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 domains for π and r are discretized with a step sizes of 0.002 and 0.05, respectively.

Panel C: Instability Tests

DGP S1							
		κ_P^I			C_P^I		
P	$R:$	50	100	200	50	100	200
50		0.04	0.04	0.04	0.05	0.05	0.05
100		0.05	0.04	0.04	0.05	0.04	0.05
200		0.05	0.05	0.04	0.05	0.05	0.04
500		0.06	0.05	0.05	0.05	0.05	0.05
1000		0.05	0.05	0.06	0.05	0.05	0.06
DGP S2							
		κ_P^I			C_P^I		
P	$R:$	50	100	200	50	100	200
50		0.04	0.04	0.04	0.05	0.05	0.05
100		0.04	0.05	0.04	0.04	0.05	0.05
200		0.05	0.05	0.05	0.05	0.05	0.05
500		0.05	0.06	0.05	0.05	0.05	0.05
1000		0.05	0.05	0.05	0.05	0.05	0.05
DGP S3							
		κ_P^I			C_P^I		
P	$R:$	50	100	200	50	100	200
50		0.04	0.04	0.05	0.05	0.05	0.05
100		0.05	0.04	0.05	0.05	0.05	0.05
200		0.05	0.05	0.04	0.05	0.05	0.05
500		0.05	0.06	0.05	0.05	0.05	0.05
1000		0.05	0.05	0.06	0.04	0.05	0.05

Note. The table reports empirical rejection frequencies for the test statistics κ_P^I and C_P^I in eqs. (4.1) and (4.1) at the 5% nominal size for various values of P and R. The number of Monte Carlo replications is 5,000. The domains for π and r are discretized with a step sizes of 0.002 and 0.05, respectively.

Table 3. Size Properties - Serially Correlated Case

DGP S4 - Asymptotic Critical Values with HAC Estimates							
			κ_P			C_P	
P	$R:$	50	100	200	50	100	200
50		-	-	-	-	-	-
100		0.07	0.10	0.07	0.04	0.06	0.05
200		0.08	0.11	0.08	0.04	0.06	0.04
500		0.07	0.10	0.07	0.04	0.05	0.04
1000		0.07	0.08	0.08	0.04	0.04	0.05

DGP S4 - Bootstrapped Critical Values							
		κ_P			C_P		
P	$R:$	50	100	200	50	100	200
50		0.06	0.06	0.06	0.06	0.06	0.06
100		0.03	0.05	0.08	0.03	0.05	0.07
200		0.06	0.06	0.05	0.06	0.06	0.06
500		0.07	0.08	0.05	0.08	0.07	0.05
1000		0.06	0.05	0.04	0.06	0.06	0.04

Note. The table reports empirical rejection frequencies for the test statistics κ_P and C_P in eqs. (13) and (14) at the 5% nominal size for various values of P and R. The number of Monte Carlo replications is 5,000. The domains for π and r are discretized with a step sizes of 0.002 and 0.05, respectively.

Table 4. Power Properties

	Forecast Density Tests Robust to Instabilities		Correct Specification Tests		Instability Tests	
Panel A: DGP P1						
c	κ_P	C_P	κ_P^{CS}	C_P^{CS}	κ_P^I	C_P^I
0	0.06	0.06	0.06	0.06	0.05	0.05
0.10	0.33	0.43	0.35	0.43	0.05	0.04
0.15	0.77	0.92	0.80	0.92	0.05	0.03
0.20	0.99	1.00	0.99	1.00	0.05	0.03
0.25	1.00	1.00	1.00	1.00	0.05	0.03
Panel B: DGP P2						
c	κ_P	C_P	κ_P^{CS}	C_P^{CS}	κ_P^I	C_P^I
0	0.05	0.06	0.05	0.06	0.05	0.04
0.010	0.07	0.07	0.05	0.06	0.22	0.25
0.015	0.11	0.09	0.05	0.05	0.50	0.55
0.020	0.16	0.13	0.06	0.06	0.68	0.75
0.025	0.26	0.19	0.05	0.06	0.87	0.91
0.030	0.45	0.32	0.05	0.06	0.96	0.97
Panel C: DGP P3						
T_1/T	κ_P	C_P	κ_P^{CS}	C_P^{CS}	κ_P^I	C_P^I
0	0.05	0.06	0.05	0.06	0.05	0.04
0.075	0.08	0.07	0.08	0.07	0.10	0.08
0.100	0.14	0.11	0.11	0.10	0.17	0.15
0.125	0.23	0.19	0.18	0.15	0.33	0.29
0.150	0.62	0.41	0.40	0.30	0.85	0.66
0.200	0.97	0.82	0.72	0.60	1.00	0.98

Note. The table reports empirical rejection frequencies for the test statistics under the alternatives of DGP P1, DGP P2, and DGP P3. $R = 40$, $P = 960$. The number of Monte Carlo replications is 5,000. The domains for π and r are discretized with a step sizes of 0.002 and 0.05, respectively.

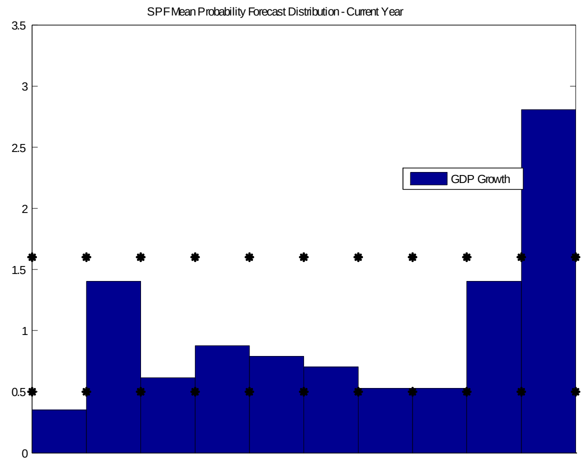
Table 5: SPF's Mean Probability Forecast Distribution

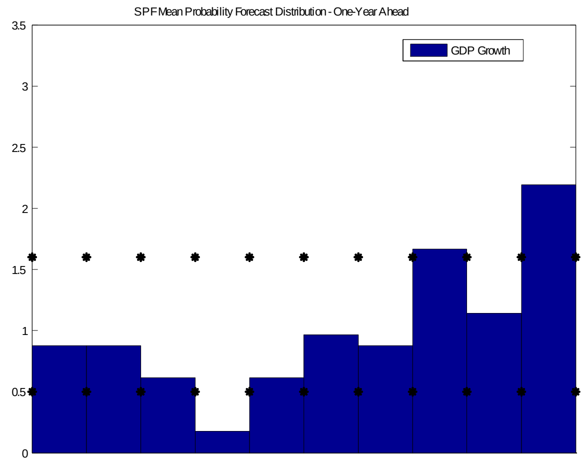
Series Name:	GDP Growth			GDP Deflator Growth		
	Forecast Density Tests Robust to Instabilities					
Horizon:	κ_P	C_P	Break	κ_P	C_P	Break
0	6.5825* †	1.9815*†	1991:IV	7.3468*†	1.8374*†	1985:III
1	4.6394* †	1.8545*†	1990:IV	2.9405*†	0.9304*†	1998:II
	Correct Specification Tests					
	κ_P^{CS}	C_P^{CS}		κ_P^{CS}	C_P^{CS}	
0	5.6326* †	1.8398*†	-	7.2455*†	1.8055*†	-
1	4.0699* †	1.7509*†	-	2.4640*†	0.7659*†	-
	Instability Tests					
	κ_P^I	C_P^I		κ_P^I	C_P^I	
0	1.1625* †	0.1417*†	1991:III	0.4221†	0.0318†	1998:IV
1	0.8020* †	0.1035*†	1992:IV	1.3129*†	0.1645*†	2001:II

Note. '*' indicates rejections at 5% significance levels under independence, while '†' indicates rejections at 5% significance levels under serial correlation respectively. The critical values under independence are as in Table 1, Panel A, while the critical values under serial correlation is simulated conditional to the data implied variance-covariance matrix for the PIT. The domains for π and r are discretized with a step sizes of 0.002 and 0.05, respectively.

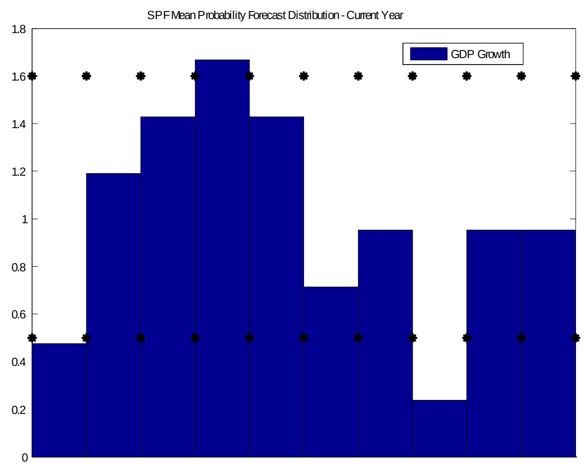
Figure 1. SPF Mean Probability Forecast Distribution - GDP Growth

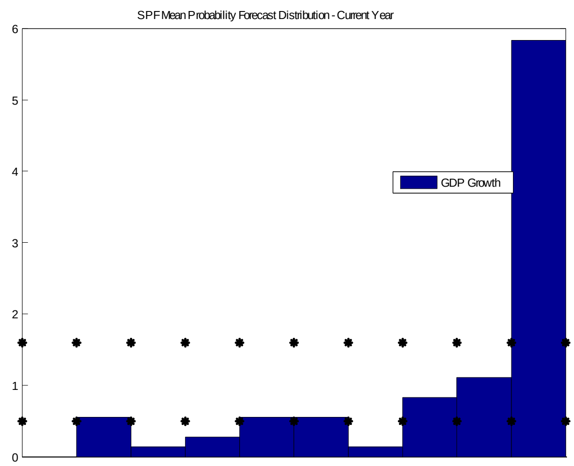
Panel A: Full Sample (1981:III-2009:IV)



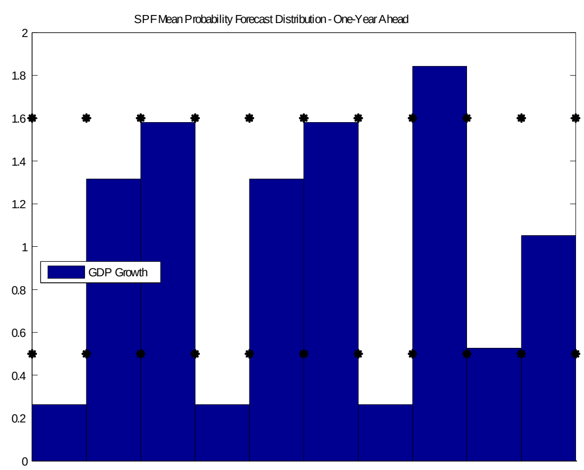


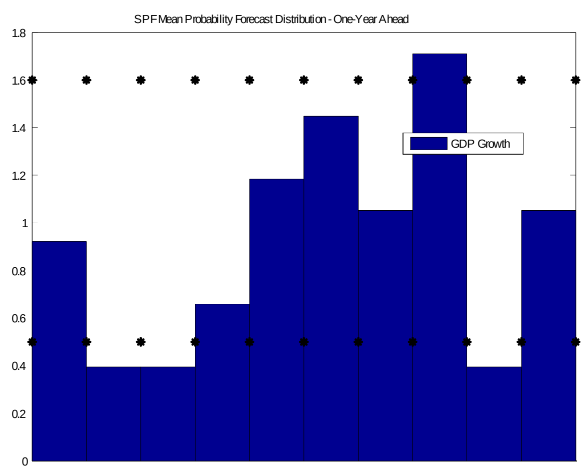
Panel B: Sub-sample Analysis for Nowcast (1981:III-1991:IV, 1992:I-2009:IV)





Panel C: Sub-sample Analysis for One-Year Ahead Forecast (1981:I-1990:IV, 1991:I-2009:IV)

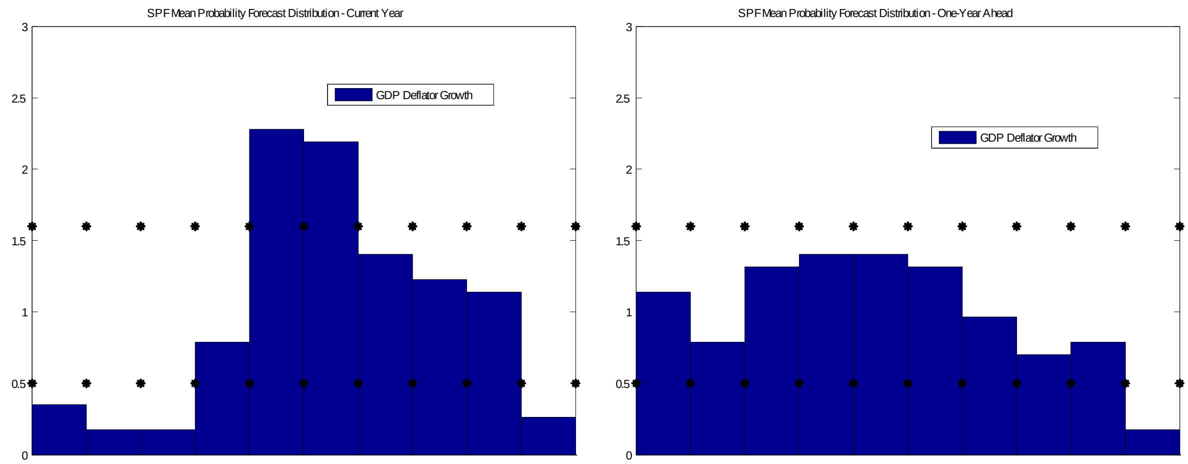




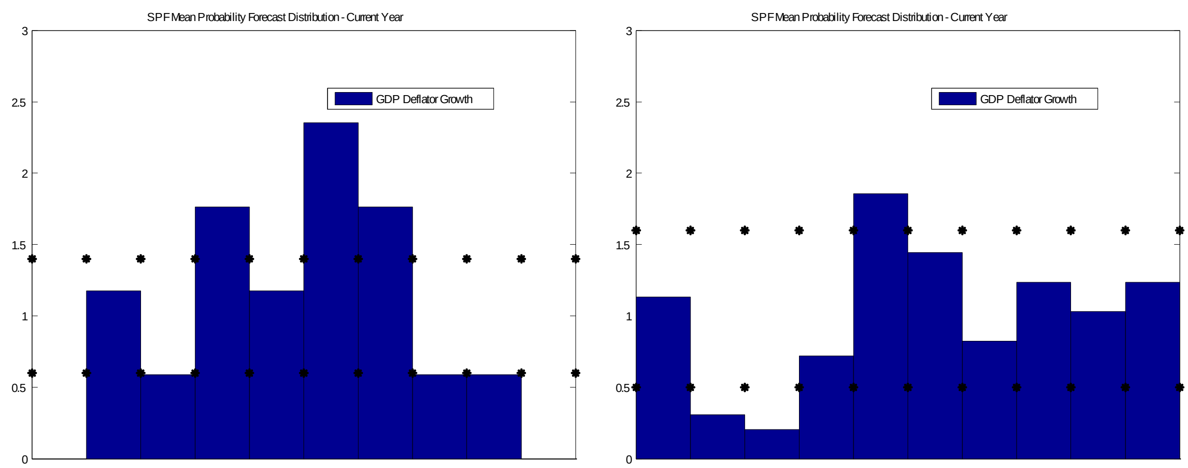
Note: The figure shows the normalized decile counts of the PITs and the 95% confidence intervals approximated under a binomial distribution.

Figure 2. SPF Mean Probability Forecast Distribution - GDP Deflator Growth

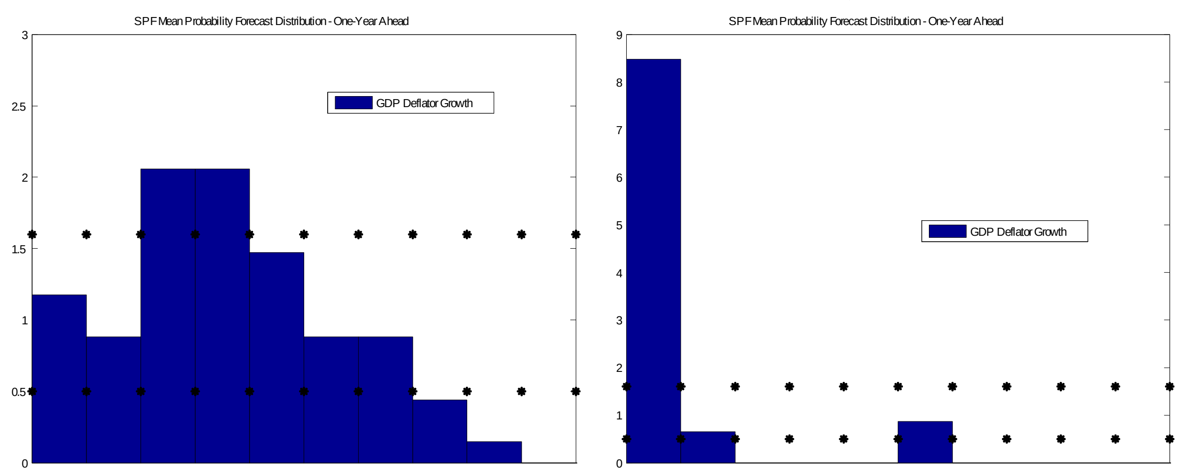
Panel A: Full Sample (1981:III-2009:IV)



Panel B: Sub-sample Analysis for Nowcast (1981:III-1985:III, 1985:IV-2009:IV)



Panel C: Sub-sample Analysis for One-Year Ahead Forecast (1981:III-1998:II, 1998:III-2009:IV)



Note: The figure shows the normalized decile counts of the PITs and the 95% confidence intervals approximated under a binomial distribution.