CENTRAL BANK & FINANCIAL SERVICES AUTHORITY OF IRELAND

EUROSYSTEM

5/RT/10 **April 2010**

Research Technical Paper

Are Some Forecasters Really Better Than Others?*

Antonello D'Agostino CBFSAI[†]

Kieran McQuinn CBFSAI[‡]

Karl Whelan UCD[§]

Economic Analysis and Research Department
Central Bank and Financial Services Authority of Ireland
P.O. Box 559, Dame Street
Dublin 2
Ireland
http://www.centralbank.ie

*The views expressed in this paper are those of the authors and do not necessarily reflect those of the Central Bank and Financial Services Authority of Ireland or the European Central Bank.

†Contact: Central Bank and Financial Services Authority of Ireland - Economic Analysis and Research Department, PO Box 559 - Dame Street, Dublin 2, Ireland. E-mail: antonello.dagostino@centralbank.ie.

‡Contact: Central Bank and Financial Services Authority of Ireland - Economic Analysis and Research Department, PO Box 559 - Dame Street, Dublin 2, Ireland. E-mail: kieran.mcquinn@centralbank.ie.

§Contact: Department of Economics, University College Dublin, Belfield, Dublin 4; E-mail: karl.whelan@ucd.ie.

Abstract

In any dataset with individual forecasts of economic variables, some forecasters will perform better than others. However, it is possible that these *ex post* differences reflect sampling variation and thus overstate the *ex ante* differences between forecasters. In this paper, we present a simple test of the null hypothesis that all forecasters in the US Survey of Professional Forecasters have equal ability. We construct a test statistic that reflects both the relative and absolute performance of the forecaster and use bootstrap techniques to compare the empirical results with the equivalents obtained under the null hypothesis of equal forecaster ability. Results suggests limited evidence for the idea that the best forecasters are actually innately better than others, though there is evidence that a relatively small group of forecasters perform very poorly.

JEL classification: C53, E27, E37.

Keywords: Forecasting, Bootstrap.

Non Technical Summary

This paper proposes a new test for assessing whether performance differences among forecasters reflect innate differences in forecasting ability and applies the test to data from the Survey of Professional Forecasters. We calculate a distribution of the performance of individual forecasters—based on a new measure of forecasting performance that combines the relative performance of the forecaster with the absolute scale of their errors—and compare these distributions with the outcomes that would have been obtained had the actual forecasts been randomly reassigned to different forecasters each period.

Based on forecasts for output and inflation over the period 1968 to 2009, our results suggest there is limited evidence for the idea that some forecasters are innately better than others, i.e. that there is a small number of really good forecasters. A sizeable minority are, however, found to be significantly worse than the bootstrap estimate. Simulations show that the presence of this underperforming group tends to result in a rather flattering appraisal of forecasters at the upper end of the performance scale. However, once the sample is restricted to exclude the worst-performing quintile, there is very limited evidence for some forecasters significantly outperforming the rest.

1. Introduction

How people formulate expectations of economic variables is one of the key methodological issues in macroeconomics. It is hardly surprising, then, there is a relatively large literature related to surveys of professional forecasters. Advocates of rational expectations have often emphasised that for the economy to behave in a fashion that is roughly compatible with rational expectations, all that is required is for agents to observe the forecasts of a small number of professionals who are incentivized to produce rational unbiased forecasts.[1] Whether such forecasters do indeed deliver such unbiased forecasts has been the subject of a number of important empirical papers such as Keane and Runkle (1992) and Bonham and Cohen (2001).

The importance of this debate about rational expectations probably accounts for the fact that most of the literature on the properties of individual-level forecasts has focused on testing for rationality and unbiasedness. There has been very little focus however on the *accuracy* of these forecasts or how this accuracy may differ across forecasters. For instance, if two individuals are both forecasting the series y_t and one produces a set of forecasts $y_t + \epsilon_{1t}$ while the other produces a set of forecasts $y_t + \epsilon_{2t}$ where both ϵ_{1t} and ϵ_{2t} are drawn from zero mean distributions, then both of these individuals are providing unbiased forecasts. However, it ϵ_{1t} is drawn from a distribution with a smaller variance than ϵ_{2t} then it is clear that the first forecaster is doing a better job than the second. If significant variations of this kind exist across forecasters, then this should have implications for how those involved in macroeconomic policy formulation should use data sets such as the Survey of Professional Forecasters and also for the public in relation to how they should process such information.

In reality, of course, we do not get to observe individuals drawing forecasts from fixed and known *ex ante* statistical distributions. All we can see are the *ex post* forecasts that individuals have provided. For this reason, the assessment of individual forecaster performance must deal explicitly with sampling variation. Casual inspection over a number of periods may reveal certain forecasters tending to reside in the upper tail of the distribution, while others appear in the lower part. However, this will not tell us whether these performances are relatively good (or relatively bad) in a statistically significant sense relative to a null hypothesis in which all individuals are drawing

[1] Once one factors in costs of gathering information, however, there are limits to how far this argument can be taken, as discussed in the classic paper of Grossman and Stiglitz (1980).

their forecasts from the same distribution.

Our paper applies a bootstrap approach to assess the extent to which the observed data on the performance of participants in the Survey of Professional Forecasters is consistent with the hypothesis of equal underlying forecasting ability. Specifically, we simulate distributions of forecast errors under the assumption of equal underlying forecast ability and compare the simulated distributions of a measure of cumulative performance with the actual outcome. The approach we take is similar to that used in research such as Kosowski, Timmerman, Wermers, and White (2006), Fama and French (2010) and Cuthbertson, Nitzsche, and O'Sullivan (2008) to assess the relative performance of mutual funds.

To our knowledge, there is only a small existing literature that addresses this question of whether some forecasters are innately better than others, with Stekler (1987) and Batchelor (1990) presenting evidence based on the Blue Chip survey and Christensen, Diebold, Rudebusch, and Strasser (2008) presenting evidence based on the Survey of Professional Forecasters. Relative to this literature, the approach taken in our paper has a number of advantages.

First, our bootstrap approach does not require a balanced panel so our paper contrasts with previous work in using all the available information on individual forecasting performance. For example, Stekler and Batchelor presented evidence based on a small sample of twenty four forecasting groups predicting GNP over the period 1977-1982. Like Christensen et al, we use data on the forecasts of individuals who participated in the Philadelphia Fed's Survey of Professional Forecasters. However, whereas Christensen et al only study three individual forecasters, our paper examines the forecasting performance of over three hundred forecasters who provided an average of twenty forecasts each.

Second, the method used by Stekler and Batchelor ascribed a rank each period to each forecaster and then summed the ranks over a number of periods to arrive at a test statistic that was used to assess the null hypothesis that the forecasters did not differ significantly in their underlying ability. This approach does not take into account the *absolute* size of any of the errors made by a forecaster, so a forecaster making the biggest error in a particular period is treated the same whatever the size of this error. In contrast, our approach is based on a test statistic for perfomance evaluation that takes into account both absolute error of the forecaster each period as well as their perfomance relative to other forecasters.

Third, rather than being a simple yes or no test of equal forecaster performance, our approach provides a graphical comparison of the realized distribution of forecaster outcomes against the distribution consistent with this null hypothesis. Our results show that the key way in which the null hypothesis of equal forecasting ability fails to hold is that there appears to be a relatively small fraction of particularly bad forecasters. Once this bottom tail is removed, there is relatively little evidence for superior ability among the remaining forecasters.

2. Testing for Differences in Forecaster Performance

This section outlines the previous work on assessing the significance of differences in forecaster performance and then describes our methodology.

2.1. Previous Work

Stekler (1987) studied the forecasts of organisations that participated in the monthly Blue Chip survey of economic indicators between 1977 and 1982. Thirty one different organisations provided forecasts but only twenty four provided forecasts for every period and his study restricted itself to studying this smaller sample. Stekler's approach assigns a score, R_{ijt} to the ith forecaster in predicting the jth variable in period t. This ranking procedure is repeated for each period under consideration. For each variable, the forecaster's scores are then summed over the whole sample of size N to produce a rank sum of

$$S_{ij} = \sum_{i=1}^{N} R_{ijt}. \tag{1}$$

Under the null hypothesis of equal forecasting ability, then each individual should have an expected rank sum score of $\frac{N(K+1)}{2}$ where K is the number of forecasters. Batchelor (1990) pointed out that, under this null, the expected rank sum has a variance of $\frac{NK(K+1)}{12}$, so the test statistic

$$g = 12 \sum_{i=1}^{K} \frac{\left(S_i - \frac{N(K+1)}{2}\right)^2}{NK(K+1)} \tag{2}$$

follows a χ^2_K distribution. Batchelor showed that the results obtained in Stekler's paper for forecasts of real GDP and inflation were not above the ten percent critical value for rejecting the hypothesis that all forecasts are drawn from the same underlying distribution.[2] Thus, for these 24 forecasting groups over this relatively short period, the evidence could be interpreted as consistent with the null hypothesis of equal forecasting ability.

Christensen, Diebold, Rudebusch and Strasser (2008) is principally a methodological paper that develops a new approach to testing for equal forecasting accuracy, extending the well-known forecast comparison test of Diebold and Mariano (1995) to a case in which there are more than two forecasts to be compared. As this method requires balanced panels and long time series for forecasts, their application to the Survey of Professional Forecasters compares the three individual forecasters who have participated most often in the survey, giving them a time series of sixty observations for each forecaster. They obtain mixed results with tests suggesting equal predictive accuracy for some variables and not others.

2.2. A Bootstrap Test

We will first describe the statistic we use to assess forecaster performance and then move on to describing our bootstrap exercise. In relation to assessing forecaster performance, the rank sum approach used by Stekler and Batchelor has a number of weaknesses. It requires a balanced panel of forecasters, which in reality is difficult to obtain because participants in forecast surveys tend to move in and out over time, so most of the information available from surveys is lost. The sum of period-by-period ranks is also likely to provide a flawed measure of forecast performance. A forecaster who occasionally does well but sometimes delivers dramatically bad forecasts may score quite well on this measure but, in reality, there may not be much demand for the professional services of someone prone to making terrible errors.

An alternative approach would be to compare forecasts according to mean square error. However, it is well known that underlying nature of macroeconomic fluctuations has changed over time. We show below that forecasting was more difficult during the period prior to the so-called Great Moderation, i.e. prior to 1984. In addition, since forecasters tend to base their projections on similar sets of publicly available information, there is a substantial common element across the forecasters. Since we are examining an unbalanced panel, we want to be careful not to attribute

[2] Stekler's paper had used an incorrect formulae for the variance for the g statistic.

superior forecasting performance to someone lucky enough to live through low-variance times.

We address these issues by measuring forecaster performance as follows. For each type of forecast that we track, we denote by N_t the number of individuals providing a forecast in period t, while the realised error of individual i is denoted as e_{it}. Because some periods are easier to forecast than others, we construct a normalised squared error statistic for each period for each forecaster defined as

$$E_{it} = \frac{{e_{it}}^2}{\left(\sum_{i=1}^{N_t} {e_{it}}^2\right)\frac{1}{N_t}} \tag{3}$$

This statistic controls for differences over time in the performance of all forecasters—each period there is a common element that can lead most forecasters to be too high or too low in their forecast—while still allowing the magnitude of the individual error to matter. For instance, an E_{it} of 2 would imply that the squared error for individual i was twice the mean squared error for that period. This method of accounting for errors does not punish forecasters simply because they contributed forecasts during unpredictable periods. However, the size of an individual's error relative to the average error for that period is taken into account.

Once these period-by-period normalised square errors have been calculated, we then assign each forecaster an overall score based on taking an average of their normalised squared error statistics across all the forecasts that they submitted. For a forecaster who first appears in the sample in period $t = TS$ and last appears in the sample in period $t = TE$, this score is

$$S_i = \frac{1}{TE - TS + 1} \sum_{j=0}^{TE-TS+1} E_{i,TE+j} \tag{4}$$

Our approach to testing the hypothesis of equal forecaster ability can be summarised as follows. Suppose that each period's forecasts were taken from the participants and were then randomly shuffled and re-assigned back to the survey participants. Would the realised historical distribution of forecaster performance—as measured by the S_i statistics—be significantly different from those obtained from this random re-shuffling? If not, then we cannot reject the hypothesis of equal underlying forecaster ability.

We apply our bootstrap technique in a way that exactly mimics the unbalanced nature of the panel we are using (the Philadelphia Fed Survey of Professional Forecasters.) Thus, corresponding to the true Forecaster 3, who joined the SPF survey in 1968:Q4 and stayed in the sample up to 1979:Q4, our bootstrapped distributions also contain a Forecaster 3 who joined and left at the same times. However, in our simulations, the forecast errors corresponding to each period are randomly re-assigned across forecasters within that period. In other words, our bootstrap simulations can be thought of as a re-running of history so that, for example, they contain a period called 1970:Q2, in which the set of forecasts actually handed in that period are randomly assigned to our simulated forecasters.[3] We do not reassign errors across periods, so our simulated forecasters for 1970:Q2 cannot be randomly assigned a forecast error corresponding to some other period.

Once we have assigned errors for each period, we calculate overall scores for each simulated forecaster using equation (4) and save the resulting distribution of scores. We then repeat this process 1,000 times, so that we have 1,000 simulated distributions, each based on randomly reassigning the errors corresponding to each period. This allows us to calculate the percentiles associated with each point in the distribution under the null hypothesis of equal forecaster ability.

For example, suppose we want to assess the outcome achieved by the best-performing forecaster. We can compare his or her outcome with both the median "best performer" from our 1,000 draws, i.e. the "typical" best performer from a random reassignment distribution. We can also compare their performance with the 5th and 95th percentiles, which give us an indication of the range that may be observed in "best performer" scores under random reassignment. If the best performer in the actual data is truly significantly better than his or her peers, we would expect their score to lie outside the range represented by these bootstrap percentiles.

3. Application to the Survey of Professional Forecasters

The Survey of Professional Forecasters (SPF) provides the most comprehensive database available to assess forecaster performance. It began in 1968 as a survey conducted by the American Statistical Association and the National Bureau for Economic Research and was taken over by the Federal Reserve Bank of Philadelphia in 1990. Participants in the SPF are drawn primarily from business

[3]The results below do this re-assignment with replacement, so that the each forecaster is assigned a forecast drawn from the same full distribution and the same individual forecast can be assigned twice. Results are essentially identical when we assign the errors without replacement.

with the survey being conducted around the middle of each quarter.

In our analysis we look at the quarterly predictions for output and its deflator.[4] We construct forecast errors for two horizons: $h = 1$, which corresponds to a "nowcast" for the current quarter and $h = 5$, which corresponds to the one year ahead forecast error. Output and inflation data are continuously revised and thus for each quarter several measures of both variables are available. Following Romer and Romer (2000), we construct the errors using the figures that were published two quarters following the date being forecasted. In other words, we assume that the aim of participants was to forecast the variable according to the measurement conventions that prevailed when the forecast was being collected.

The measure of output is Gross National Product (GNP) until 1991 and Gross Domestic Product (GDP) from 1992 onwards. The evaluation sample begins in 1968:Q4 and ends in 2009:Q3. In total N = 309 forecasters appear in the survey over the time period and the average amount of time spent in the sample is five years or twenty forecasts.

Figure 1 provides an illustration of the raw data used in our analysis. It shows the forecast errors for the nowcast of inflation and output over the entire sample (1968 - 2009) with lines of different colours corresponding to different individual forecasters. The figure illustrates two aspects of forecasting that we noted earlier.

First, the figure makes it clear that for most periods, there is a significant common element across forecasters in their errors, so that for some quarters almost all errors are positive while for other periods almost all are negative. The importance of this common component explains why our measure of perfomance normalises the individual squared errors by the average squared error for that period. Second, the significant reduction in variation in the forecast errors from the mid-1980s onwards, which corresponds with the "great moderation", is notable. This result has been commented upon before by Stock and Watson (2005, 2006) and D'Agostino, Giannone and Surico (2006) amongst others from a forecasting perspective. In our analysis, we assess the robustness of our findings by performing our analysis on pre- and post-moderation samples as well as the full sample.

[4]The data used are taken from the website of the Federal Reserve Bank of Philadelphia.

4. Results

We present our results in two ways, graphically and in tables.

4.1. Results for All Forecasters

Table 1 provides the results from applying our method to the full sample of 309 forecasters. The figures in the rows of the table are the scores corresponding to various percentiles of the empirical distribution of forecasting performance for our four types of forecasts (GDP current quarter and next year, inflation over the current quarter and over the next year). The figures in brackets correspond to the fifth and ninety-fifth percentiles generated from our bootstrap distributions.

Table 1 can be read as follows. Taking the figures in the first row, 0.249 is the score obtained by the forecaster who was placed at the fifth percentile in projecting current quarter GDP i.e. the forecaster who performed better than 95 percent of other forecasters. The figures underneath (0.156-0.326) correspond to the fifth and ninety-fifth percentiles of the 1000 simulated scores for forecasters who placed in this position. In other words, five percent of our bootstrap simulations produced fifth percentile scores less than 0.156 and five percent produced fifth percentile scores greater than 0.326 (since these are average normalised square errors, low scores indicate a good performance). Because the realized first-percentile score of 0.249 fits comfortably in between these two figures, we can conclude that the actual fifth percentile forecasters of current quarter GDP were not statistically significantly different from what would be obtained under a distribution consistent with equal underlying ability.

More generally, the results from this table show that scores of the top performing forecasters—those in the upper fifth percentiles for forecasting current quarter inflation as well as year-ahead forecasts for GDP and inflation—are generally well inside the ninety fifth percentile bootstrap intervals generated from random reassignment. The middle percentiles of the empirical distribution have scores that are lower than the bootstrap distribution (implying lower errors for these percentiles than generated under the null of equal underlying ability). Because the average scores from the realised and bootstrap distributions are the same by construction, these are offset by scores for the poorer forecasters that are higher than generated by the bootstrap distributions.

This pattern is not well picked up by the specific percentiles reported in Table 1 but can be understood better from Figure 2. This figure shows the cumulative distribution function (CDF) from the SPF data (the dark line) along with the fifth, median, and ninety-fifth bootstrap percentiles for each position in the distribution (the thin lines). The empirical CDF generally stays close to these bootstrap distributions, with the main deviations being somewhat lower scores in the middle of the empirical distribution being offset by somewhat higher scores for some of the weakest performers. (These patterns are a bit hard to see for current quarter forecasts for inflation because the scores for some of the poor performers are so big relative to the majority of other participants.)

4.2. Results for Restricted Samples of Forecasters

One potential problem with these results is that they treat all forecasters equally, whether they contributed two forecasts and then left the SPF panel or whether they stayed in the panel for ten years. Thus, some of the "best" forecasters—both in the data and in our bootstrap simulations—are people (either real or imagined) who participated in a small number of surveys and got lucky. So, for example, the best performing forecaster for current quarter inflation has a normalised average square error of 0.000; similarly, the fifth bootstrap percentiles for best forecasters are also zero. To reduce the influence of those forecasters who participated in a small number of editions of the survey, we repeat our exercise excluding all forecasters who provided less than ten forecasts. Thus, we restrict our attention to those who have participated in the survey for at least two and a half years.

Table 2 and Figure 3 provide the results from this exercise. In relation to the best forecasters, the results here are mixed. The best forecasters for current quarter inflation and year-ahead GDP are significantly better than those generated by the bootstrap simulations while the best forecasters for current quarter GDP and year-ahead inflation are not. However, beyond the very top of the distribution, the forecasters in the top half of the distribution generally all have scores that are superior to those generated from the bootstrapping exercise. That said, what emerges most clearly from Figure 3 is that these significantly low scores are offset by a relatively small number of very bad performances that are far worse than predicted by the bootstrap distributions. In other words, the empirical distribution differs mainly from those generated under the null hypothesis of equal forecaster performance in having a small number of very bad forecasters.

This result provides an answer to the question posed in our title. Some forecasters really are better than others. However, a better way to phrase this result is that some forecasters really are worse than others. This raises a final question: If we excluded those forecasters who clearly performed badly, can we find evidence that there are significant differences among the rest. To get at the answer to this question, we re-run our bootstrapping exercise, still excluding those with less than ten forecasts but this time also excluding those forecasters who scored worse than the eightieth percentile. These results are presented in Table 3 and Figure 4.

We draw two principal conclusions from these results. First, in relation to the best forecasters in the SPF, these performances are not statistically different relative to the upper ends of the distributions generated from the bootstrap exercise based on randomly reassigning the forecasts from this best eightieth percent of forecasters. Second, looking at Figure 4, the empirical distributions for GDP and inflation at both horizons are, at almost all points in the distribution, very close to the bootstrap distributions.

The principal conclusion that we draw from these results is that apart from the strong evidence that there is some forecasters who perform very poorly in the SPF, perhaps because they do not take participation in the survey very seriously, there is limited evidence for innate differences between the remaining eighty percent or so of participating forecasters.

4.3. Pre- and Post-1985 Samples

As a final exercise, we performed our analysis using samples restricted to the pre- and post-moderation, which we date here as 1985. It may be that the nature of forecasting changed significantly with the onset of this moderation, so it may be worth checking whether these two periods generate very different results. Figures 5 and 9 show the data for individual forecast errors from these two periods, while Figures 6-8 and Figures 10-12 replicate Figures 2-4 for these separate two samples.

While there are some differences the general flavour of the results are pretty similar across the two time periods. The unrestricted distributions (including all forecasters, Figures 6 and 10) are very similar to the bootstrap distributions, particularly for those with low average error scores. When attention is restricted to those with ten or more forecasts (Figures 7 and 11) there is some evidence that the better performers have lower scores than generated by the bootstrap distributions, particularly for inflation. However, these deviations are mainly accounted for by the very poor perfomances of a small number of bad forecasters. When attention is restricted to the best 80 percent of forecasters (Figures 8 and 12) the shape of the actual distributions are generally very close to those generated by the bootstrap with random reassignment.

5. Conclusions

This paper has presented a new test for assessing whether performance differences among forecasters reflect innate differences in forecasting ability and applies the test to data from the Survey of Professional Forecasters. We calculated a distribution of the performance of individual forecasters—based on a new measure of forecasting performance that combines the relative performance of the forecaster with the absolute scale of their errors—and compared these distributions with the outcomes that would have been obtained had the actual forecasts been randomly reassigned to different forecasters each period.

Based on forecasts for output and inflation over the period 1968 to 2009, our results suggest there is only limited evidence for the idea that some forecasters are innately better than others, i.e. that there is a small number of really good forecasters. A sizeable minority are, however, found to be significantly worse than the bootstrap estimate. Simulations show that the presence of this underperforming group tends to result in a rather flattering appraisal of forecasters at the upper end of the performance scale. However, once the sample is restricted to exclude the worst-performing quintile, there is very limited evidence for some forecasters significantly outperforming the rest.

On balance, we conclude that most of the participants in the Survey of Professional Forecasters appear to have approximately equal forecasting ability.

References

[1] Bonham, Carl and Richard Cohen (2001), "To Aggregate, Pool, or Neither: Testing the Rational Expectations Hypothesis Using Survey Data," *Journal of Business and Economic Statistics*, 19, 278291.

[2] Batchelor, Roy A. (1990), "All Forecasters are Equal." *Journal of Business and Economic Statistics* 8(1), 143 - 144.

[3] Christensen H. Jens, Francis X. Diebold, Georg H. Strasser and Glenn D. Rudebusch (2008). "Multivariate Comparison of Predictive Accuracy", working paper available at http://www.econ.uconn.edu/Seminar

[4] Cuthbertson, Keith, Dirk Nitzsche and Niall O'Sullivan (2008). "UK Mutual Fund Performance: Skill or Luck?" *Journal of Empirical Finance*, 15, 613-634.

[5] D'Agostino, Antonello, Domenico Giannone, and Paolo Surico (2006), "(Un)Predictability and macroeconomic stability," Working Paper Series 605, European Central Bank.

[6] Diebold, Francis. X. and Mariano, Roberto (1995). "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, 13, 253263.

[7] Fama, Eugene F and Kenneth French (2010). "Luck versus Skill in the Cross Section of Mutual Fund Returns," forthcoming, *Journal of Finance*.

[8] Grossman, Sanford and Joseph Stiglitz (1980). "On the Impossibility of Informationally Efficient Markets," *American Economic Review*, 70, 393-408.

[9] Keane, Michael and David Runkle (1990). Testing the Rationality of Price Forecasts: New Evidence from Panel Data, *American Economic Review*, 80, 714-735.

[10] Kosowski, Robert, Allan Timmerman, Russ Wermers and Hall White (2006). "Can Mutual Fund "Stars" Really Pick Stocks? New Evidence from a Bootstrap Analysis," *Journal of Finance*, 56, 2551-2595.

[11] Mehra, Yash. (2002), Survey Measures of Expected Inflation: Revisiting the Issue of Predictive Content and Rationality. Federal Reserve Bank of Richmond Economic Quarterly 88, 17-36.

[12] Romer, David and Christine Romer (2000) “Federal Reserve information and the behavior of interest rates”, *American Economic Review* 90, 429-457.

[13] Stekler, Herman. (1987), “Who Forecasts Better?” *Journal of Business and Economic Statistics* 5(1), 155 - 158.

[14] Stock, James and Mark Watson (2005). Has inflation become harder to forecast? Prepared for the conference “Quantitative Evidence on Price Determination”, Board of Governors of the Federal Reserve Board, September 29-30, Washington DC.

[15] Stock, James and Mark Watson (2006). Why has U.S. inflation become harder to forecast? National Bureau of Economic Research (NBER) Working paper 12324.

Table 1: Distribution of Forecasting Performance With Bootstrap 5th and 95th Percentiles

	Percentiles					
1 quarter	Best	5	25	50	75	Worst
GDP	0.016	0.249	**0.578**	**0.792**	1.170	**21.501**
	(0.000 - 0.025)	*(0.156-0.326)*	*(0.632-0.710)*	*(0.866-0.927)*	*(1.116-1.206)*	*(3.743 - 15.802)*
Inflation	0.000	0.232	**0.536**	**0.761**	**1.189**	9.622
	(0.000-0.022)	*(0.178-0.319)*	*(0.606-0.687)*	*(0.850-0.918)*	*(1.127-1.227)*	*(3.718 - 16.037)*
1 year	Best	5	25	50	75	Worst
GDP	0.016	0.316	**0.571**	**0.793**	1.154	8.758
	(0.008-0.131)	*(0.212-0.384)*	*(0.642-0.715)*	*(0.861-0.923)*	*(1.104-1.192)*	*(3.622-22.009)*
Inflation	0.033	0.359	**0.627**	**0.798**	1.143	7.615
	(0.000 - 0.058)	*(0.265-0.415)*	*(0.660-0.730)*	*(0.876-0.934)*	*(1.113-1.200)*	*(3.400 - 15.410)*

Note: The table reports the empirical distribution of forecaster performance for 309 forecasters from the SPF. The measure of forecaster performance, which is the average of the normalised squared error, E_{it} as defined in equation (3) of the paper. The figures in brackets refer to the fifth and ninety-fifth percentiles generated by the bootstrap distribution obtained under the null hypothesis of equal forecaster ability.

Table 2: Distribution of Forecasting Performance: Restricted to Those With At Least 10 Forecasts

			Percentiles			
1 quarter	Best	5	25	50	75	Worst
GDP	0.321	**0.503**	**0.655**	**0.825**	1.131	**6.742**
	(0.255 - 0.482)	*(0.531 - 0.632)*	*(0.756 - 0.817)*	*(0.921 - 0.976)*	*(1.112 - 1.191)*	*(1.957 - 3.362)*
Inflation	**0.232**	**0.458**	**0.629**	0.782	1.039	**3.728**
	(0.243 - 0.455)	*(0.560 - 0.651)*	*(0.760 - 0.822)*	*(0.919 - 0.976)*	*(1.105 - 1.182)*	*(1.916 - 3.362)*
1 year	Best	5	25	50	75	Worst
GDP	**0.321**	**0.500**	**0.635**	**0.836**	1.146	2.901
	(0.327 - 0.511)	*(0.537 - 0.632)*	*(0.744 - 0.811)*	*(0.912 - 0.972)*	(1.105 - 1.190)	(1.986 - 4.035)
Inflation	0.408	**0.500**	**0.695**	**0.883**	1.111	**4.720**
	(0.330 - 0.529)	*(0.560 - 0.651)*	*(0.760 - 0.822)*	*(0.919 - 0.976)*	(1.105 - 1.182)	*(1.916 - 3.362)*

Note: The table reports the empirical distribution of forecaster performance for the 176 forecasters who contributed at least ten quarterly forecasts to the SPF between 1968 and 2009. The measure of forecaster performance, which is the average of the normalised squared error, E_{it} as defined in equation (3) of the paper. The figures in brackets refer to the fifth and ninety-fifth percentiles generated by the bootstrap distribution obtained under the null hypothesis of equal forecaster ability.

Table 3: Distribution of Forecasting Performance: Best 80 Percent With At Least 10 Forecasts

	Percentiles					
1 quarter	Best	5	25	50	75	Worst
GDP	0.405	0.591	**0.728**	**0.935**	**1.178**	2.171
	(0.320 - 0.560)	*(0.589 - 0.693)*	*(0.805 - 0.863)*	*(0.949 - 0.997)*	*(1.100 - 1.165)*	*(1.640 - 2.538)*
Inflation	0.337	0.593	**0.751**	**0.940**	1.166	2.381
	(0.301 - 0.545)	*(0.577 - 0.685)*	*(0.800 - 0.859)*	*(0.948 - 0.997)*	*(1.103 - 1.170)*	*(1.666 - 2.598)*
1 year	Best	5	25	50	75	Worst
GDP	0.436	0.641	**0.795**	**0.944**	**1.156**	1.952
	(0.417 - 0.617)	*(0.624 - 0.719)*	*(0.813 - 0.870)*	*(0.946 - 0.995)*	*(1.088 - 1.155)*	*(1.605 - 2.476)*
Inflation	0.438	**0.595**	**0.806**	0.972	**1.182**	2.144
	(0.389 - 0.612)	*(0.628 - 0.724)*	*(0.821 - 0.876)*	*(0.953 - 0.999)*	*(1.092 - 1.155)*	*(1.558 - 2.347)*

Note: The table reports the empirical distribution of forecaster performance for the best-performing eighty percent of the 126 forecasters who contributed at least ten quarterly forecasts to the SPF between 1968 and 2009. The measure of forecaster performance, which is the average of the normalised squared error, E_{it} as defined in equation (3) of the paper. The figures in brackets refer to the fifth and ninety-fifth percentiles generated by the bootstrap distribution obtained under the null hypothesis of equal forecaster ability.

Figure 1: *Output and Inflation Forecast Errors*

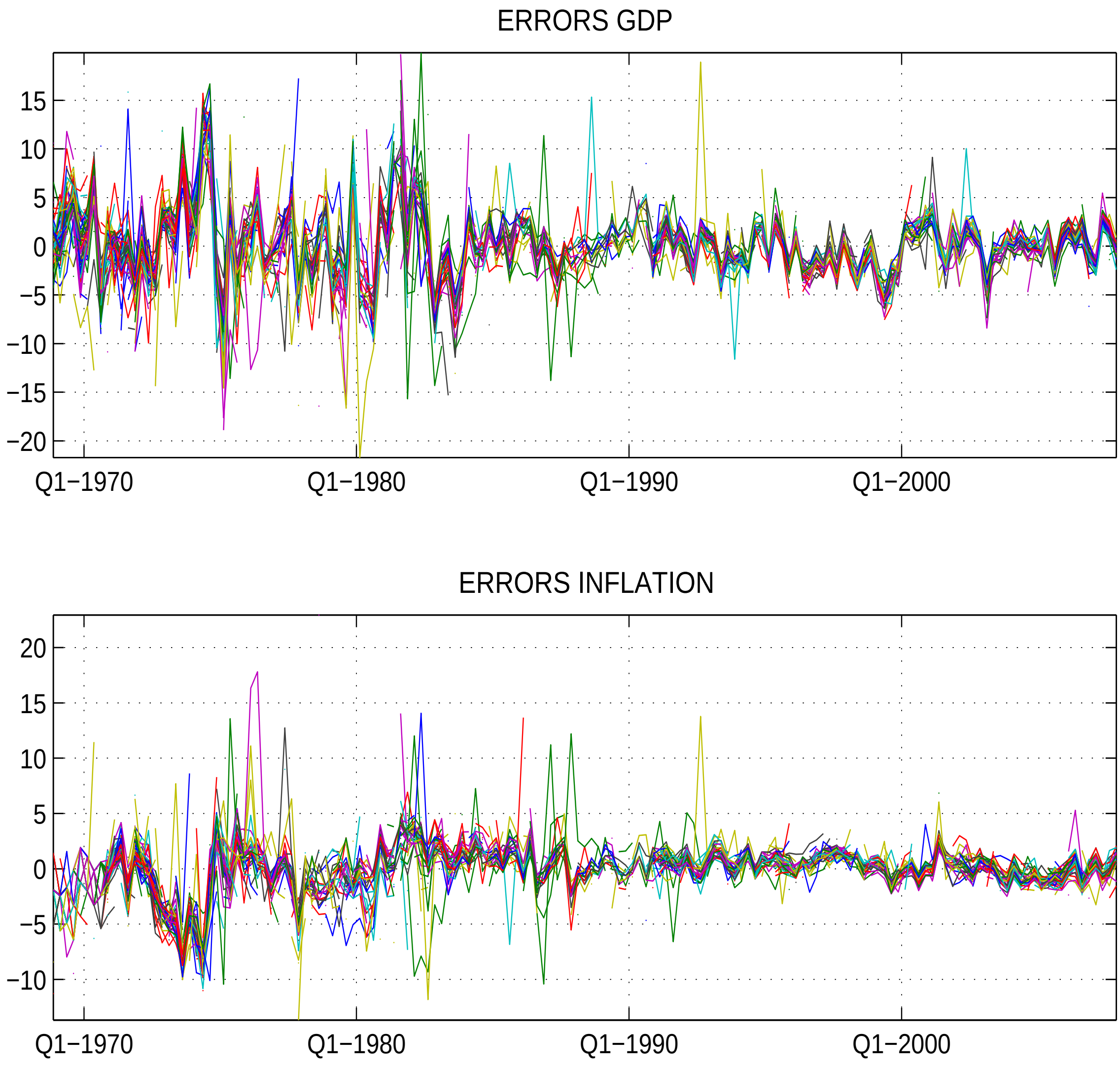


Figure 2: *Actual and Bootstrap Distributions (5th, 50th, 95th Percentiles): All Forecasters*

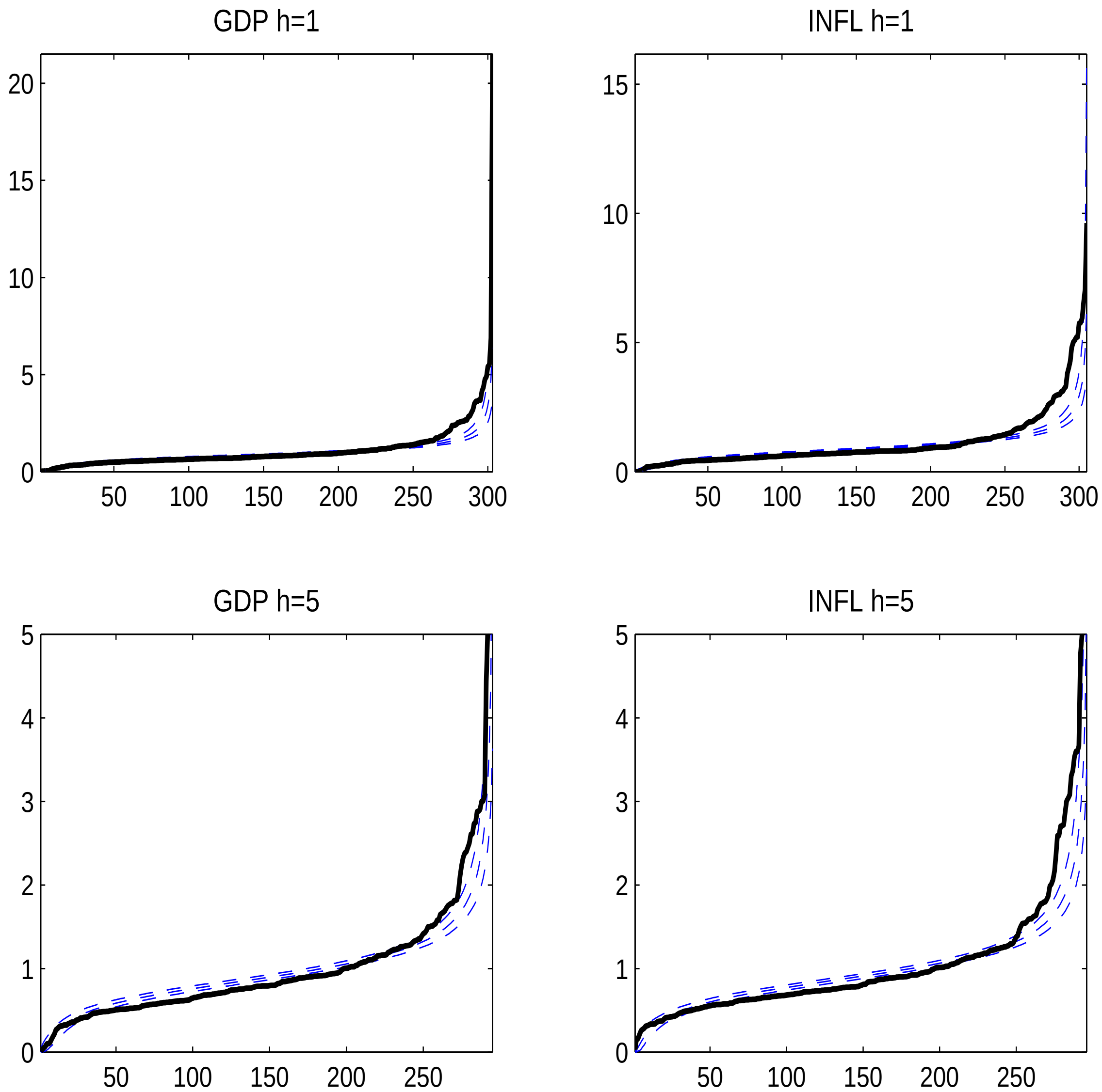


Figure 3: *Actual and Bootstrap Distributions: Minimum of Ten Forecasts*

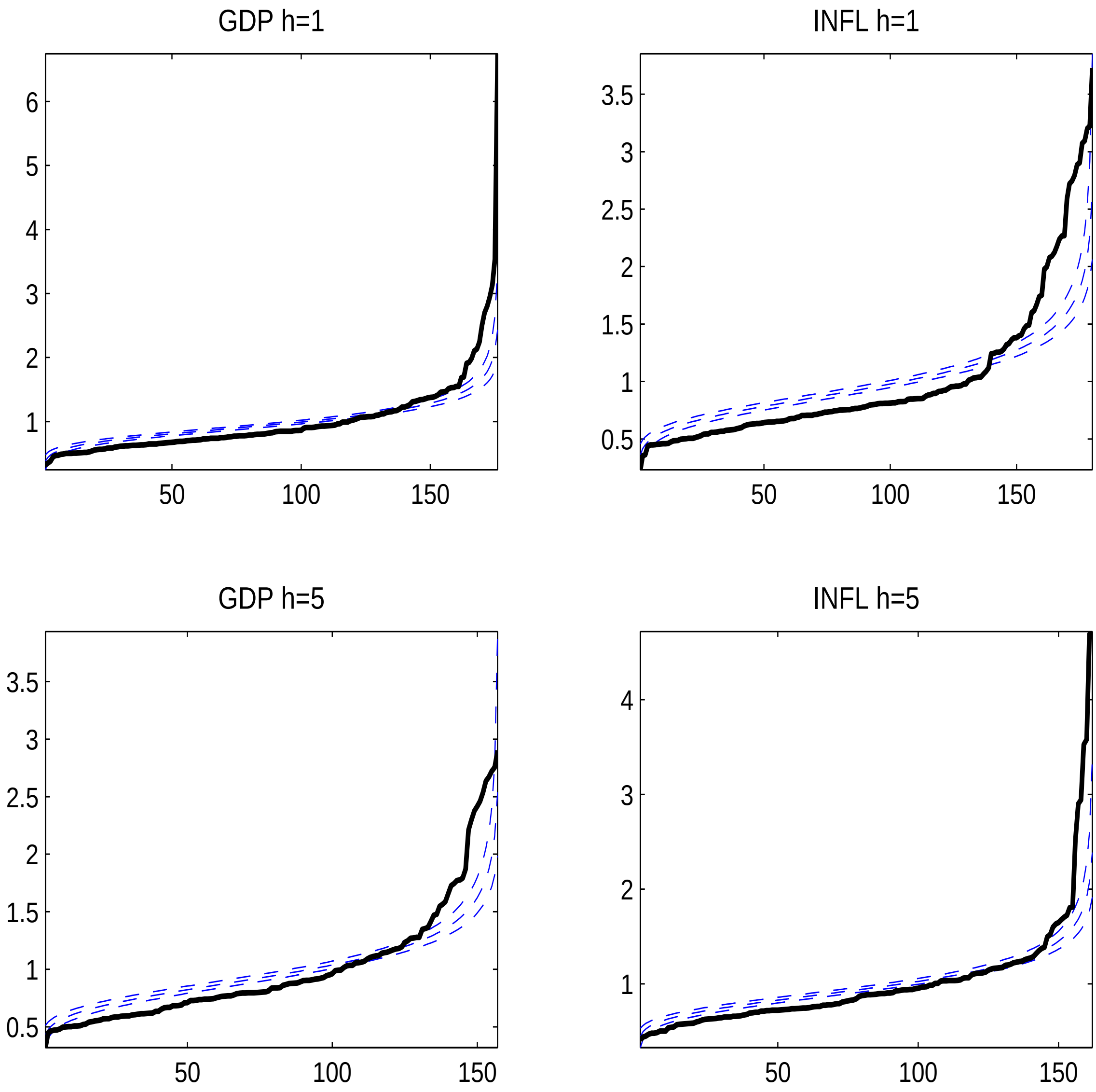
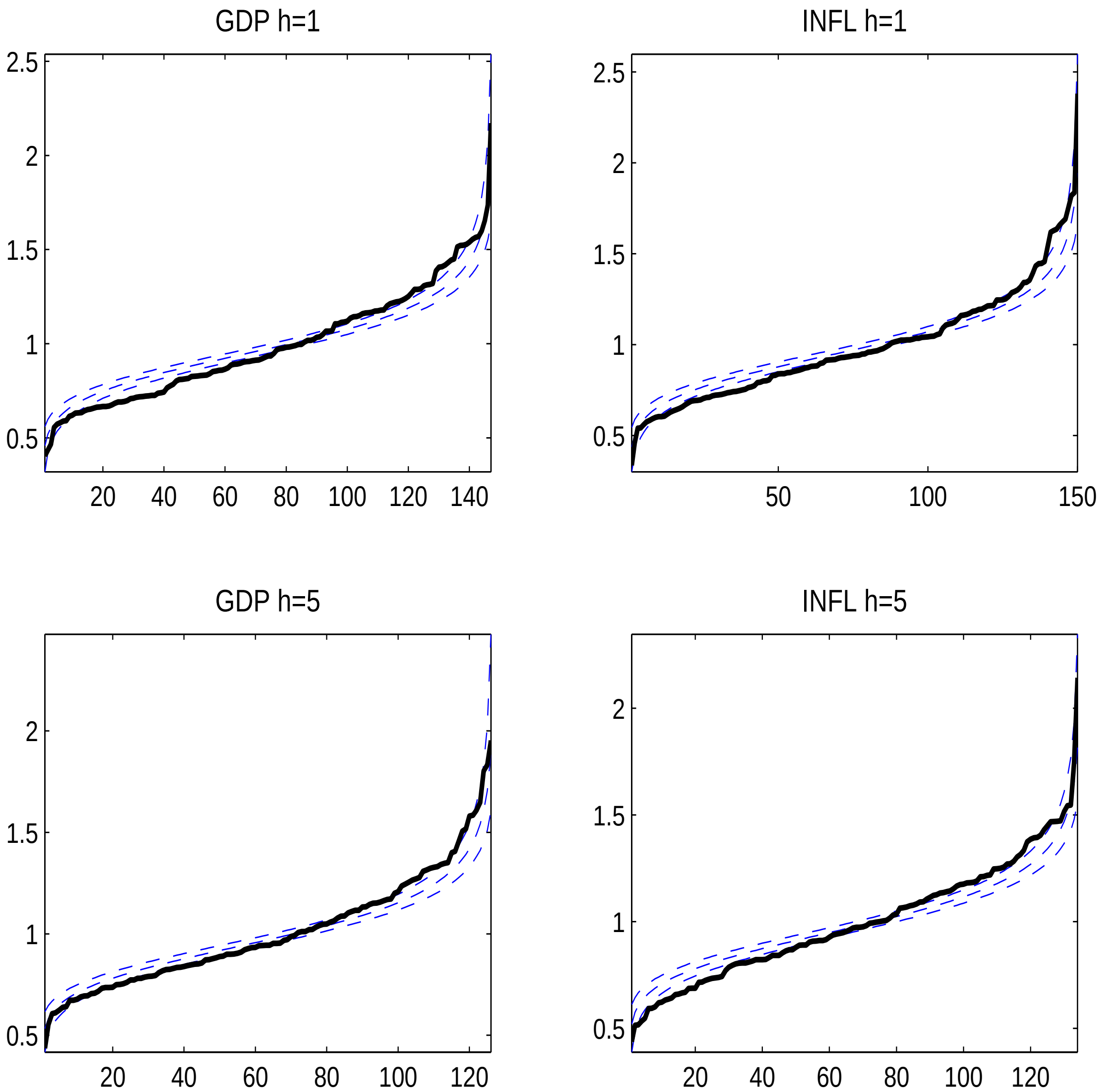
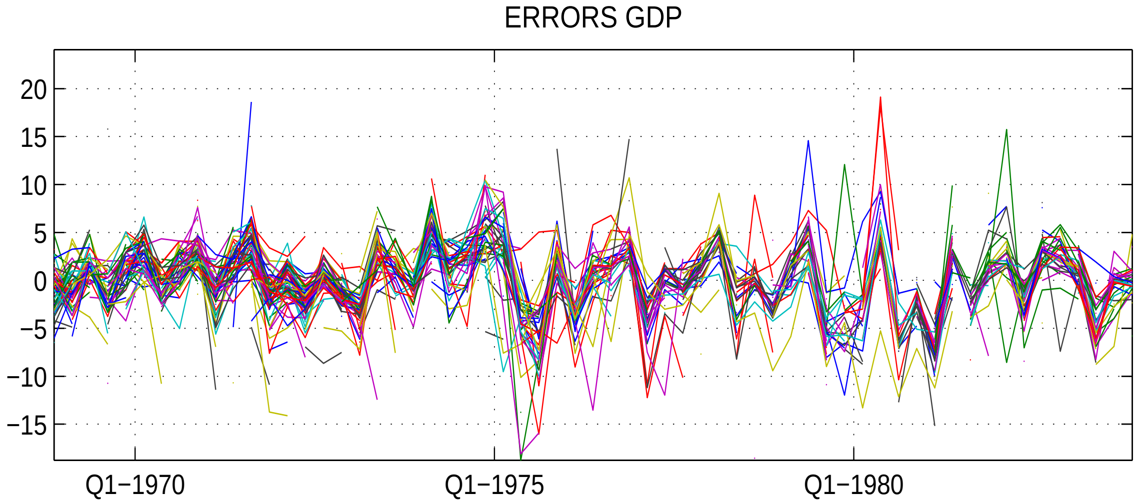


Figure 4: *Actual and Bootstrap Distributions: Minimum of Ten Forecasts (Best 80 Percent)*



Pre-85 Sample

Figure 5: *Output and Inflation Forecast Errors*



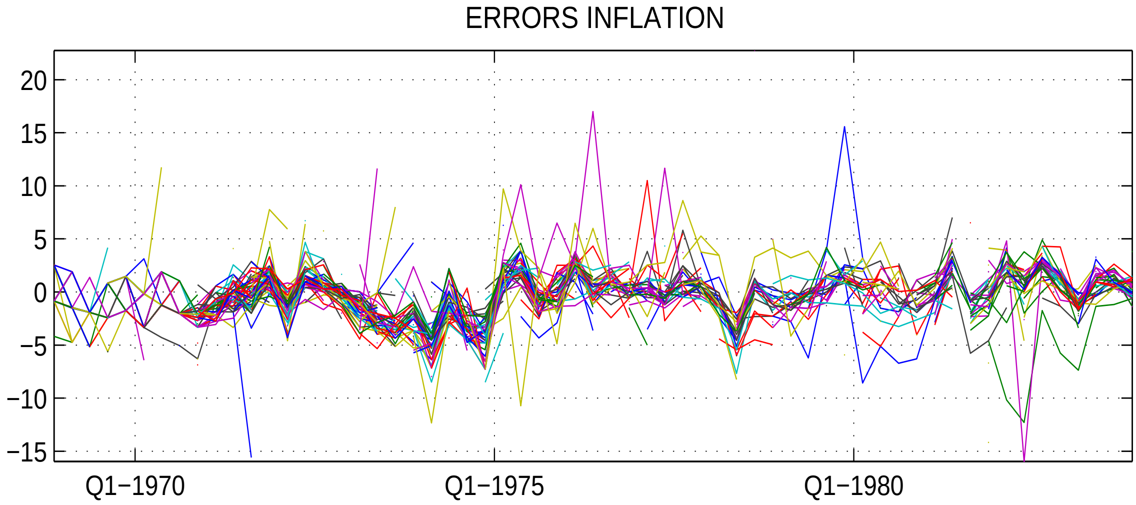


Figure 6: *Actual and Bootstrap Distributions (5th, 50th, 95th Percentiles): All Forecasters*

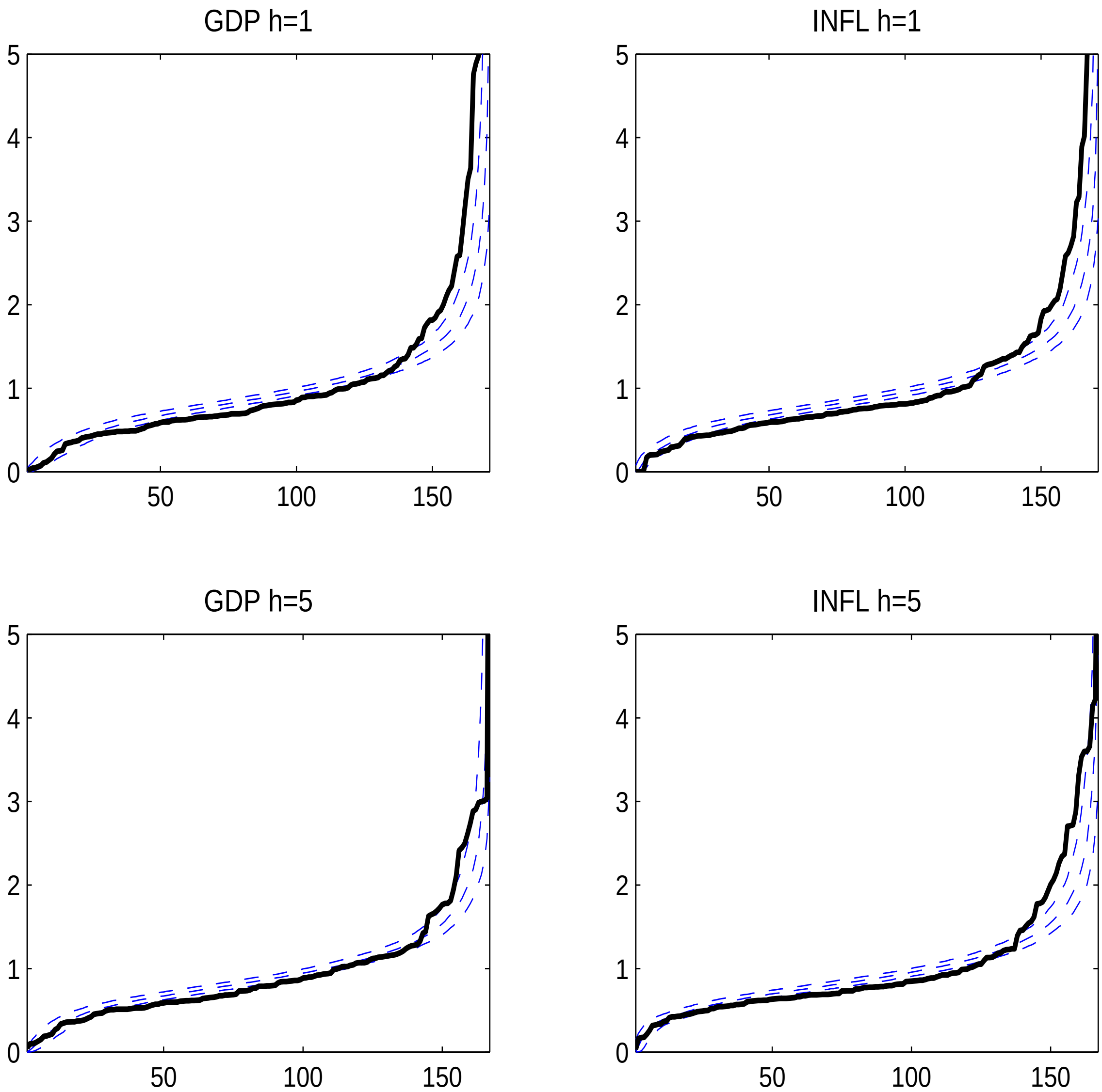


Figure 7: *Actual and Bootstrap Distributions: Minimum of Ten Forecasts*

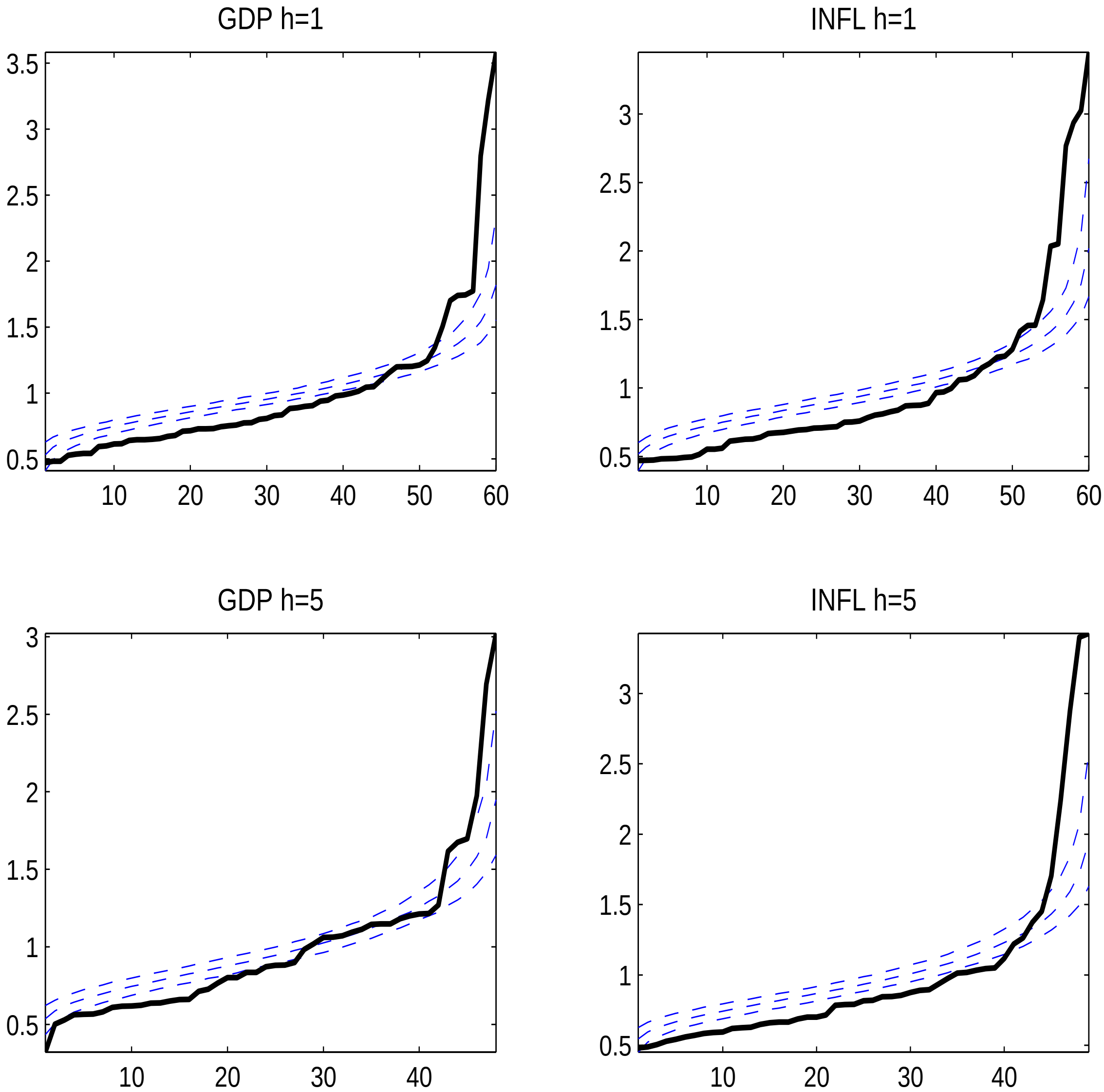
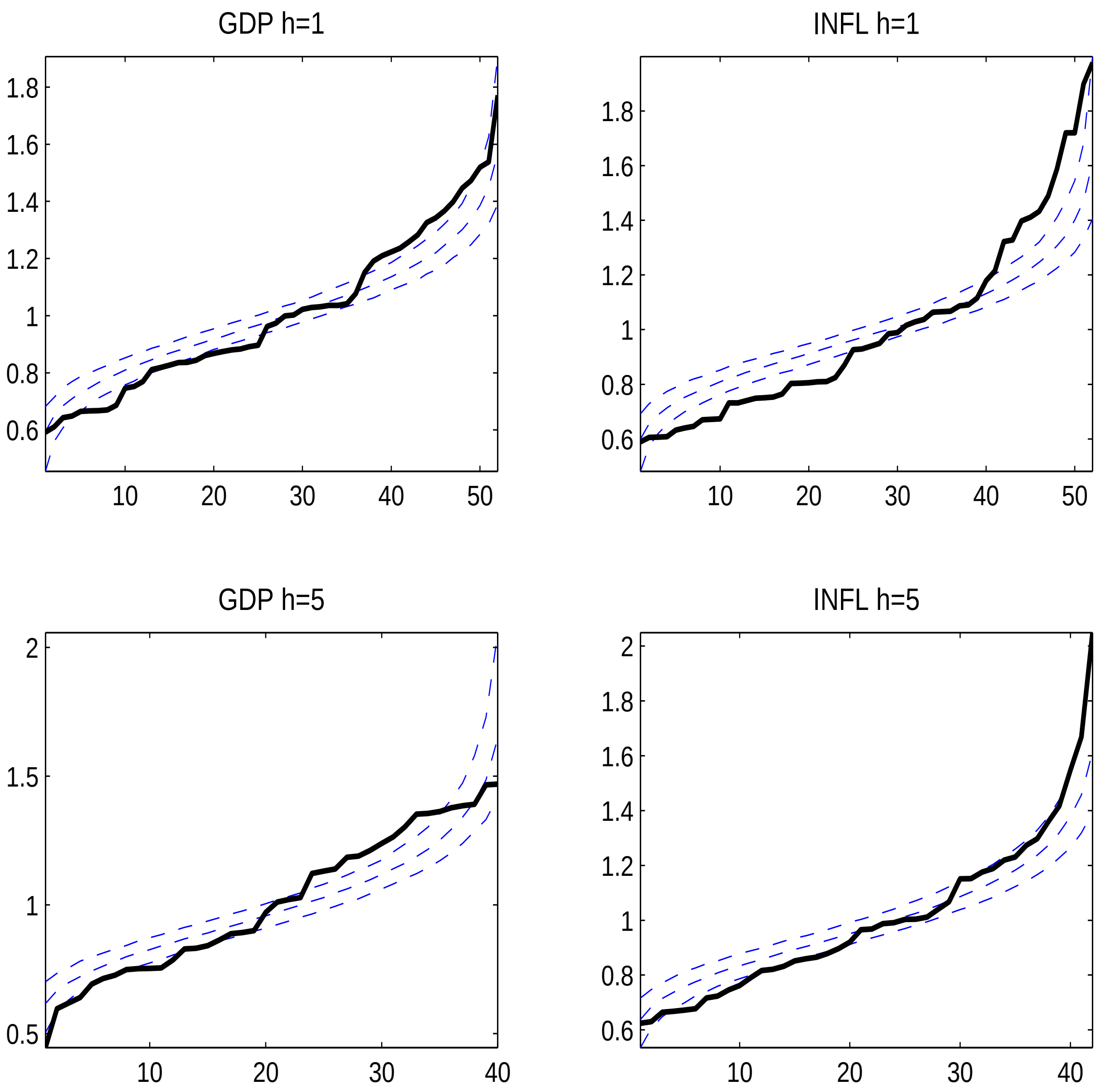


Figure 8: *Actual and Bootstrap Distributions: Minimum of Ten Forecasts (Best 80 Percent)*



Post-85 Sample

Figure 9: *Output and Inflation Forecast Errors*

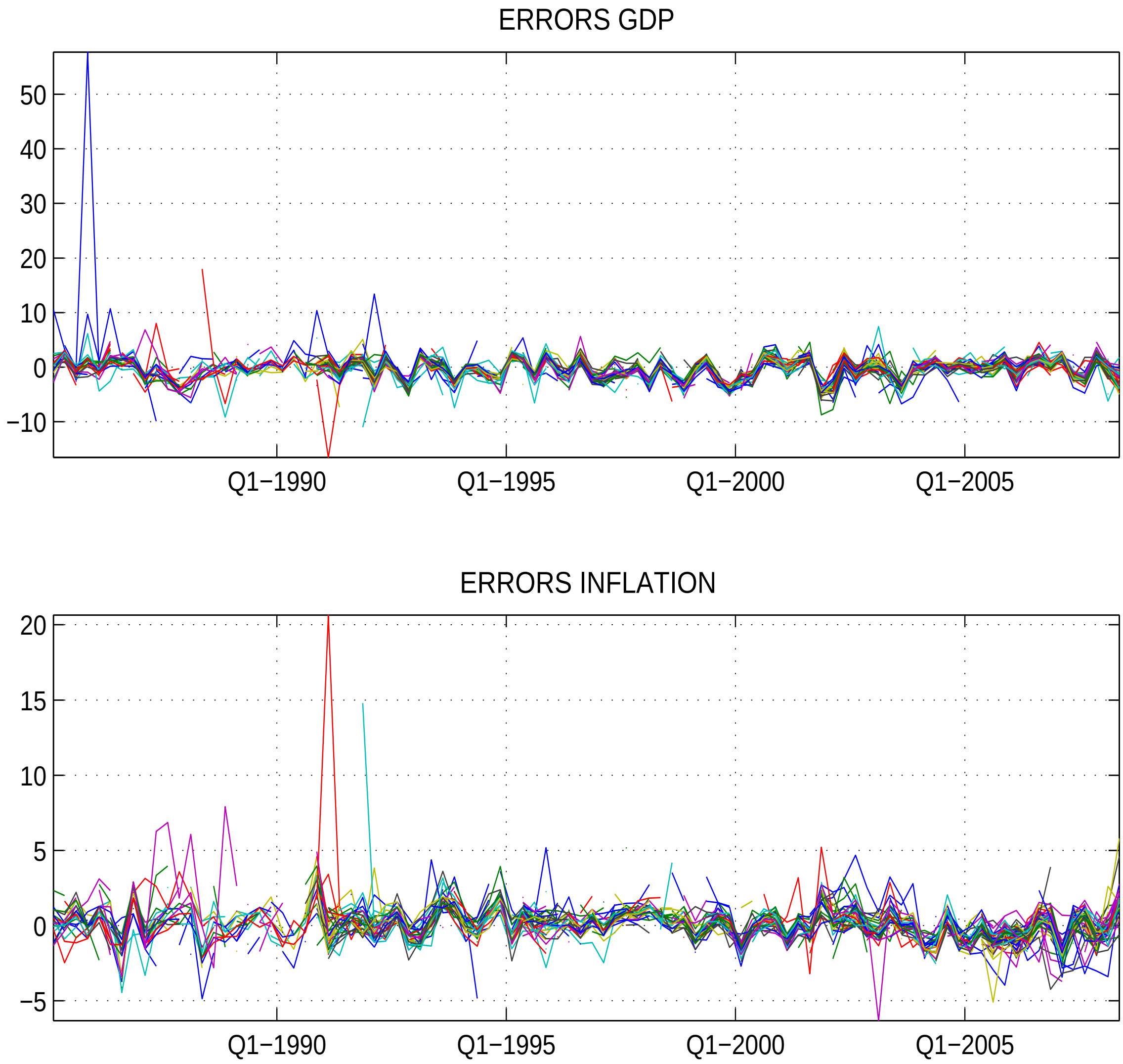


Figure 10: *Actual and Bootstrap Distributions (5th, 50th, 95th Percentiles): All Forecasters*

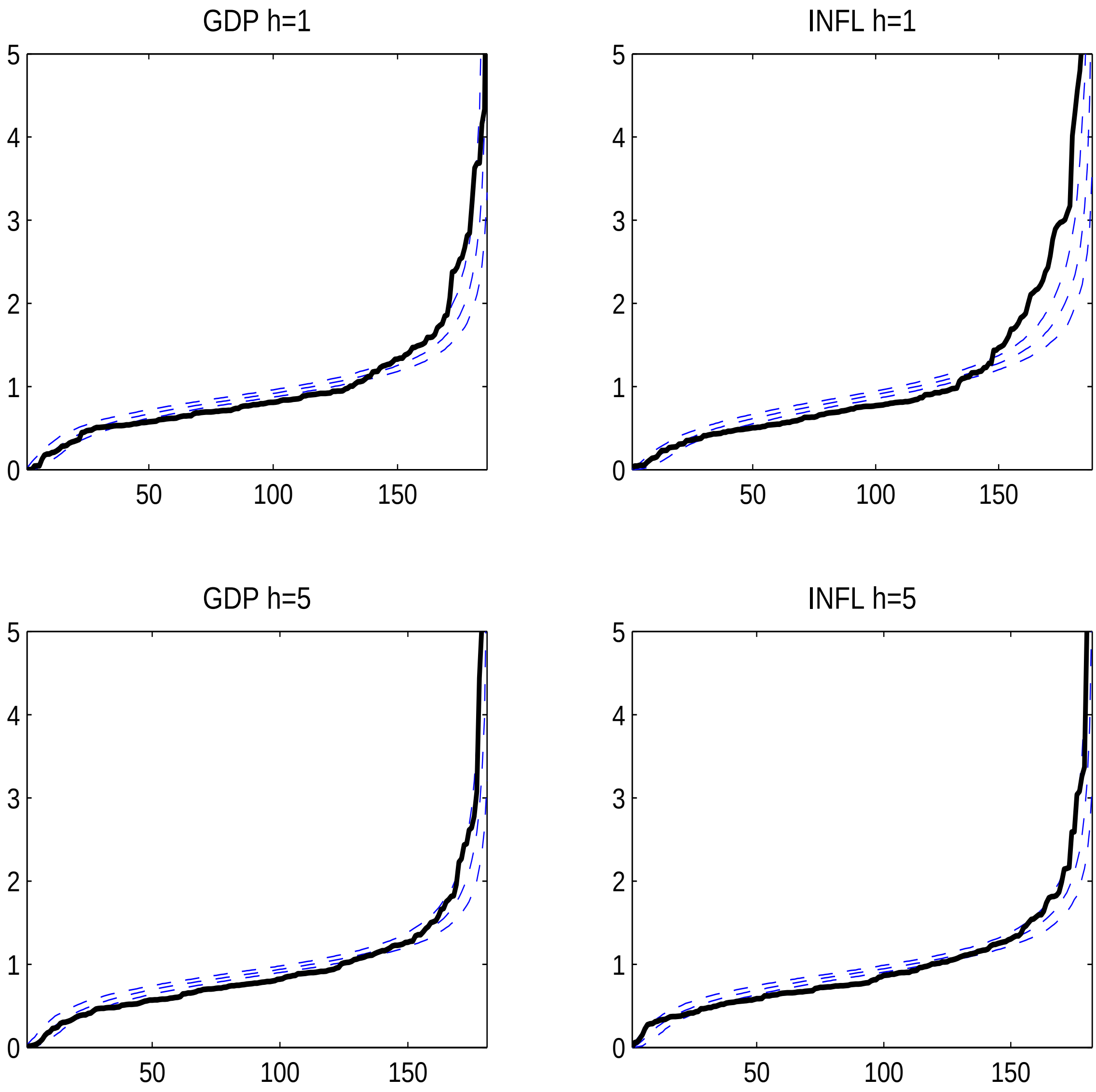


Figure 11: *Actual and Bootstrap Distributions: Minimum of Ten Forecasts (Best 80 Percent)*

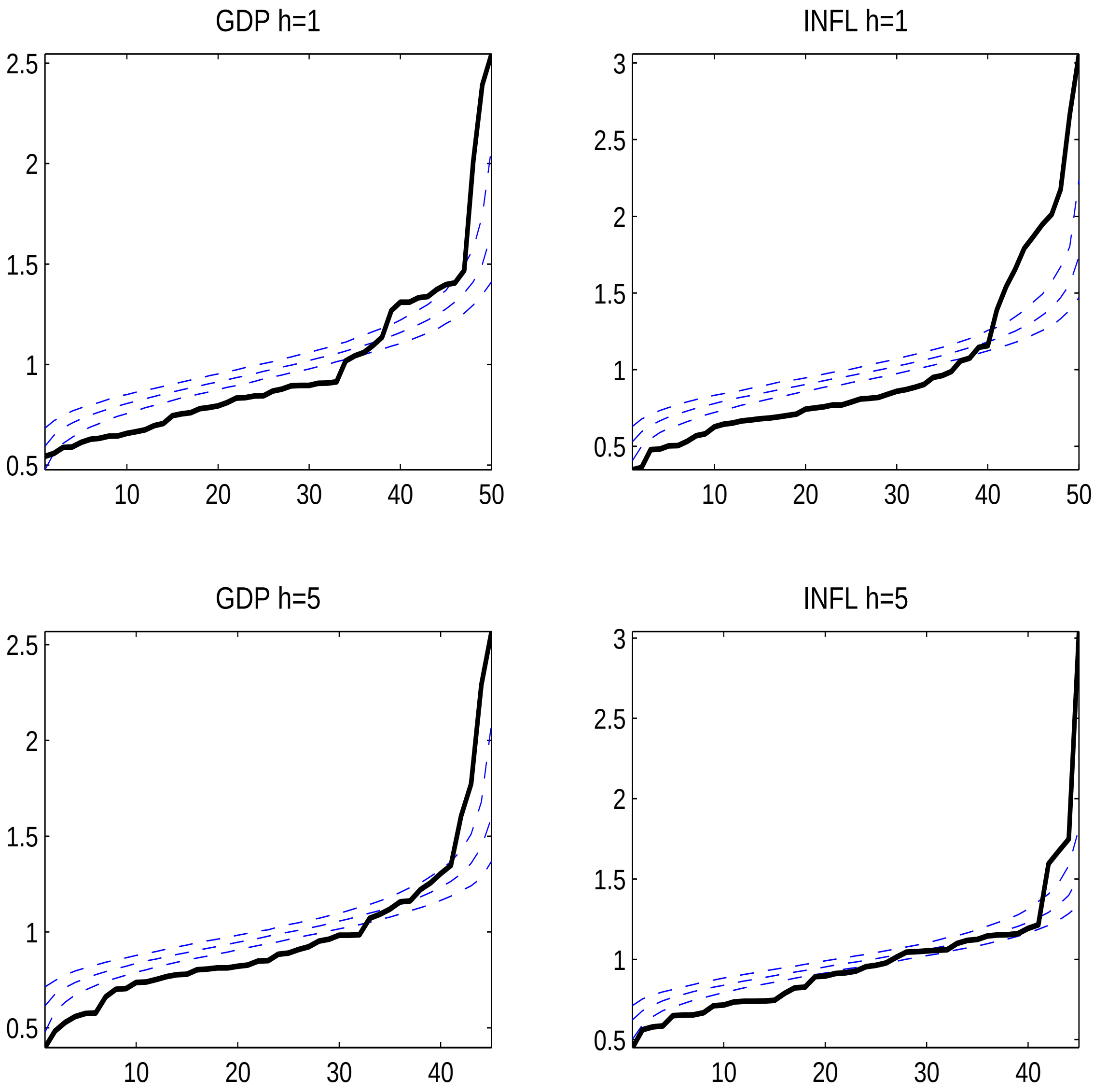


Figure 12: *Actual and Bootstrap Distributions: Minimum of Ten Forecasts (Best 80 Percent)*