JULY 2012

How Informative are the Subjective Density Forecasts of Macroeconomists?[†]

Geoff Kenny,[a,b] Thomas Kostka[a] and Federico Masera[c]

ABSTRACT

In this paper, we propose a framework to evaluate the subjective density forecasts of macroeconomists using micro data from the euro area Survey of Professional Forecasters (SPF). A key aspect of our analysis is the evaluation of the entire predictive densities, including an evaluation of the impact of density features such as location, spread, skew and tail risk on density forecast performance. Overall, we find considerable heterogeneity in the performance of the surveyed densities at the individual level. Relative to a set of simple benchmarks, this performance is somewhat better for GDP growth than for inflation, although in the former case it diminishes substantially with the forecast horizon. In addition, we report evidence of some improvement in the relative performance of expert densities during the recent period of macroeconomic volatility. However, our analysis also reveals clear evidence of overconfidence or neglected risks in expert probability assessments, as reflected in frequent occurrences of events which are assigned a zero probability. Moreover, higher moment features of expert densities, such as skew or the degree of probability mass in their tails, are shown not to contribute significantly to improvements in individual density forecast performance.

Keywords: forecast evaluation, neglected risks, real-time data, Survey of Professional Forecasters

JEL: C22, C53

[†] The authors would like to thank Malte Knüppel, Sylvain Leduc, James Mitchell, Robert Rich and Ken Wallis for useful comments and discussions as well as participants at the 2nd CesIfo conference on Macroeconomics and Survey Data in Munich, 11 and 12 November 2011, as well as participants in the Eurosystem's Working Group on Forecasting meeting in Bratislava on 22-23 September 2011. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the ECB or the Eurosystem. Any errors are the sole responsibility of the authors.

[a] European Central Bank, Kaiserstrasse 29, D-60311 Frankfurt, Germany

[b] Corresponding author: DG Research, European Central Bank, Kaiserstrasse 29, D-60311 Frankfurt, Germany. Email: Geoff.kenny@ecb.europa.eu, tel: +49 69 1344 6416, fax: +49 69 1344 6575

[c] c/o Universidad Carlos III de Madrid, Economics Department, C/ Madrid 126, 28903 Getafe (Madrid), España

Non-technical summary

Economic agents and policy makers often rely on macroeconomic risk and uncertainty assessments to inform their choices and decisions. Formally, such expert assessments are represented in the form of a complete density forecast which summarises the probability attaching to all possible future outcomes and not just a central or most likely outcome. Not much, however, is really known concerning the overall quality and accuracy of this type of information. For example, are macroeconomic experts able to provide risk and uncertainty predictions that are superior to simple rules of thumb or even very naïve statistical statements, e.g. equivalent to flipping a coin? If so, which specific features of these density forecasts contribute to strengthening their predictive power? Additionally, are macroeconomic experts better able to assess the uncertainty surrounding some economic variables, such as output growth, rather than others such as inflation? Or, do such risk assessments have equivalent "validity" or information content at both short and longer term horizons or during particular business cycle episodes, e.g. during recessions or following a financial crisis? Answers to such questions should be of interest to policy makers or anyone else who relies on expert risk assessments when making their decisions and economic choices.

In this paper, as a way of shedding some light on these questions, we propose various methods to evaluate the information contained in the density forecasts of professional macroeconomists. The surveyed forecasts we examine refer to euro area macroeconomic developments and are collected as part of the ECB Survey of Professional Forecasters (SPF). We evaluate the surveyed density forecasts in the SPF relative to various crude benchmark alternatives. The benchmarks are based on simple distributional assumptions and condition only on information that was publicly available at the time the survey was carried out. They therefore represent a somewhat low-lying threshold against which to assess the "skill" or incremental information content that may be embodied in the subjective density forecasts of macroeconomic experts. To exploit fully the information in the survey, our analysis focuses on the *individual* density forecasts. This reflects the possible bias that can result from examining aggregate forecasts alone. A key aspect of our framework is the use of evaluation measures which consider the entire predictive densities, a feature which facilitates an evaluation of the impact of density features such as their location, spread, skew and tail risk on density forecast performance. Controlling for the role of differences in point forecast accuracy (location), this analysis helps shed light on whether or not these "higher order" density characteristics also contribute to improving forecast performance.

Our study highlights a number of key features of expert density forecasts, including considerable heterogeneity in individual performance. Although a large share of experts are unable to perform better than our proposed set of simple benchmarks, a number of the better-performing experts are able to do so. The ability of some experts to improve on the mechanical benchmarks is strongest for forecasting GDP growth where, at shorter horizons of one year, 75% of all experts outperform the best performing benchmark. At longer forecasting horizons of two years, however, the superior risk assessment of the best performing experts for GDP growth is lost. Indeed, for GDP growth at this longer horizon, an agnostic uniform benchmark performs better than any of the experts in the panel. In the case of inflation forecasts, we uncover again considerable heterogeneity in the performance of expert densities. However, for this variable, only 10% of the experts are able to outperform the best performing benchmark. Indeed, several of the experts exhibit a performance that is significantly worse than our proposed crude benchmarks.

Given the inclusion in our sample of the period of macroeconomic volatility associated with the recent financial crisis, we have also examined whether the performance of expert densities changes during such key phases of the business cycle. Indeed, we uncover some scope for many experts to improve their performance during such periods. This is stronger for GDP growth densities than it is for inflation and it is only observable at relatively short horizons. However, our analysis also suggests clear evidence of overconfidence or neglected risks in the probability assessments of professional macroeconomists, as reflected in frequent occurrences of events which are assigned a zero probability. In addition, higher moment features of the surveyed expert density forecasts, such as their skew or changes in the degree of probability mass given to the tails of the predictive distributions tend - as a rule - not to contribute significantly to improvements in individual density forecast performance.

The above findings should be of interest to both producers (i.e. survey participants) as well as end users of such density forecasts. In particular, experts need to reassess the degree of confidence they have embodied in their probability assessments. Similarly, forecast users need to bear in mind the limited information value of some expert predictions and the large heterogeneity that exists in predictive performance of macroeconomic experts at the individual level. An important avenue for future research with such survey data will be to explore in more detail this heterogeneity and, in particular, whether the information content of such surveys for forecast users can be enhanced through more optimal aggregation of the constituent expert opinions.

1. INTRODUCTION

The recent financial crisis and its associated heightened levels of macroeconomic volatility, have underlined the limitation of point forecasts as a sufficient information variable for macroeconomic policy deliberations. Indeed, as surveyed in Tay and Wallis (2000), a large forecasting literature exists on the importance of density forecasts and the need to supplement forecast information on expected future values of forecast target variables with additional information from the forecast variable's entire predictive density. Moreover, in practice, it would appear that central banks and other policy institutions often rely as much – if not more - on assessments of the risks to the macroeconomic outlook, and in particular the risks to price stability, when considering alternative policy choices.[1] More broadly, during the recent financial crisis and subsequent recession, the large policy responses across the global economy, embracing substantial monetary and fiscal policy stimuli, was crucially informed by a risk analysis that highlighted the downside risks to economic activity and, in some instances, also a risk of deflation[2]. However, the primacy of *risk assessments* as a decisive source of information underpinning many policy choices extends beyond the monetary and fiscal policy domains. In the newly prioritised area of macro prudential supervision and financial stability analysis, a key aspect underpinning the assessments of vulnerabilities in the financial sector and their systemic implications includes a review and identification of relevant macroeconomic risks.[3]

In all of the above contexts, such risk and uncertainty assessments may draw on model based analysis but also include more subjective and judgemental elements or simply expert opinions. Indeed, even if model-based, some element of judgement is always embodied in such information, reflecting, for example, the choice of a particular modelling strategy or model selection criteria. While such risk assessments are evidently a key aspect underpinning both policy discussions and subsequent policy decisions and also private sector choices, not much is really known concerning the overall quality and accuracy of this information. For example, are

[1] Killian and Manganelli (2008) have related risk management concepts to the analysis of central bank preferences. Looking at US monetary policy, they find evidence against the quadratic and symmetric preferences that underpin the Taylor rule.

[2] In the case of the euro area, as an illustration of this, in explaining its' decision to reduce its key policy rates by 75 basis points on 4 December 2008, the Governing Council of the ECB noted ". . . the economic outlook remains surrounded by an exceptionally high degree of uncertainty. Risks to economic growth lie on the downside".

[3] Many definitions and concepts of risk and uncertainty are in popular use. Here, we use the terms risk and uncertainty assessments interchangeably to refer to any information from a forecaster's predictive density that describes probabilities around a central (e.g. modal) prediction. Such risk and uncertainty assessments can relate to the overall level of uncertainty, as measured by the variance of the predictive density. However, they can also refer to directional aspects as captured by a density's skew or forecasts of tail events as reflected in the magnitude of probability mass in the extremities of a given distribution.

macroeconomic experts able to provide risk and uncertainty predictions that are superior to simple rules of thumb or even very naïve statistical statements, e.g. equivalent to flipping a coin? If so, which specific features of these density forecasts contribute to strengthening their predictive power? Additionally, are macroeconomic experts better able to assess the uncertainty surrounding some economic variables, such as output growth, rather than others such as inflation? Or, do such risk assessments have equivalent "validity" or information content, at both short- and longer-term horizons or during particular business cycle episodes? Answers to such questions should be of interest to policy makers or anyone else who relies on expert risk assessments when making their decisions and economic choices.

In this paper, as a way of shedding some light on these questions, we propose and implement various methods to evaluate the information contained in the surveyed density forecasts of macroeconomists. The surveyed forecasts that we examine refer to macroeconomic developments in the euro area and are collected as part of the ECB Survey of Professional Forecasters (SPF).[4] The study of such "real-time forecasts", as collected in surveys, came to prominence with the work of Zarnowitz (1969) and Zarnowitz and Lambros (1987) and contrasts with much of the applied forecasting literature which focuses on the "pseudo" real-time predictions of specific models which have often not been available for actual decision making. In two earlier studies of SPF data, Lahiri, Teigland and Zaporowski (1988) estimate the first four moments of the individual SPF distributions and relate them to interest rates while Diebold, Gunther and Tay (1998) employ the probability integral transform to evaluate the densities. However, surveyed forecasts have been the subject of renewed attention in macroeconomics in recent times. For example, Ang, Bekaert and Wei (2007) stress the usefulness of the inflation forecasts collected in such surveys compared with other benchmarks, such as the predictions implicit in asset markets or those derived econometrically from other macroeconomic variables. Ghysels and Wright (2009) investigate the determinants of professional forecasters' forecasts while Giordani and Söderlind (2006) explore the possible role of surveyed densities in explaining the equity premium puzzle using US SPF data. Engelberg, Manski and Williams (2009) and Clements (2010) compare the point predictions of professional forecasters with their subjective probability distributions. One strand in the literature on density evaluation, as surveyed in Tay and Wallis (2000), and represented by Wallis (2003) and Dowd (2008) and Knüppel and Schulterfrankenfeld (2011), has focussed on the evaluation of the density forecasts produced and published by central

[4] Additional information about the survey as well as the complete underlying micro dataset can be downloaded from http://www.ecb.europa.eu/stats/prices/indic/forecast/html/index.en.html. See Bowles *et al.* (2010) for detailed background information related to the ECB SPF.

banks. The latter study has, for example, attempted to test empirically the optimality of measures of skew extracted from central bank density forecasts, finding that they have little systematic information content. Finally Boero, Smith and Wallis (2011) and Casillas-Olvera and Bessler (2006) have compared central bank and private sector density forecasts collected in surveys.

Using individual level data from the euro area SPF, the present paper is also intended to shed light on the ability of macroeconomists to correctly identify macroeconomic risks as reflected in their surveyed density forecasts. To conduct our evaluation, we assess the predictive performance of the surveyed density forecasts relative to various crude benchmark alternatives. The chosen benchmarks are based on simple distributional assumptions and condition only on information that was publicly available at the time the survey was carried out. They therefore represent a somewhat low-lying threshold against which to assess the "skill" or incremental information content, embodied in the subjective density forecasts of macroeconomic experts. The logic behind this approach reflects the view that to the extent that professional economists can deliver insightful information on macroeconomic risks, e.g. regarding skew or heightened probabilities of more extreme events, they should be able to outperform these simple and mechanical benchmarks. To exploit fully the information in the survey, our analysis focuses primarily on the *individual* density forecasts and not just on the distribution of point forecasts or even the combined density forecast that is often reported in official publications.[5] This reflects the possible bias that can result from examining aggregate forecasts alone – a concern which is even more relevant when one is aggregating entire distributions and not merely point forecasts.[6] To take account of this clear potential for aggregation bias, when seeking to draw inference about the ability of macroeconomists to provide informative assessments, we report separately the evaluation results for the cross-section of individual survey replies together with the results for the aggregation of individual probabilities. Finally, we also propose a set of cross-sectional and panel regressions to examine the role of key distributional features, in explaining density forecast performance both across time and across individuals. Controlling for the role of differences in point forecast accuracy (density location), this analysis helps shed light on whether or not higher

[5] Rich and Tracy (2010) consider alternative measures of uncertainty using micro data from the US SPF.

[6] Indeed, we would argue that a meaningful answer to the question posed in the title of the paper can only be based on the individual level forecasts. This reflects the fact that, as discussed in Wallis (2005), Geweke and Aminsano (2011) and Hall and Mitchell (2007), the act of aggregating densities into a linear opinion pool transforms the individual densities along most dimensions that are relevant to the question at hand. In particular, as a finite mixture distribution, the equal weighted combination of density forecasts will tend to have a variance that is larger than the average variance of its constituent densities reflecting the impact of disagreement across models or survey participants on the location of the true underlying density. The aggregate density may also exhibit other properties such as positive or negative skew and / or 'fat tails" even if the underlying individual densities do not exhibit such features.

order density characteristics, such as variance, skew or the fatness of a density forecast's tails also contribute to improving forecast performance.

The layout of the remainder of the paper is as follows. In Section 2 we describe in detail the evaluation framework we employ. Section 3 provides some summary information on distributional properties of the surveyed density forecasts we examine. Section 4 presents the main empirical results evaluating the forecast performance of the density forecasts for GDP growth and HICP inflation at both 1 and 2-year ahead horizons. Section 5 uses both cross section and panel regressions to quantify and test the role of distributional features (location, spread, skew and fatness of tails) in explaining density forecast performance. Finally, Section 6 concludes with our overall assessment of the ability of macroeconomists to forecast risks as reflected in the relative performance of their surveyed density forecasts.

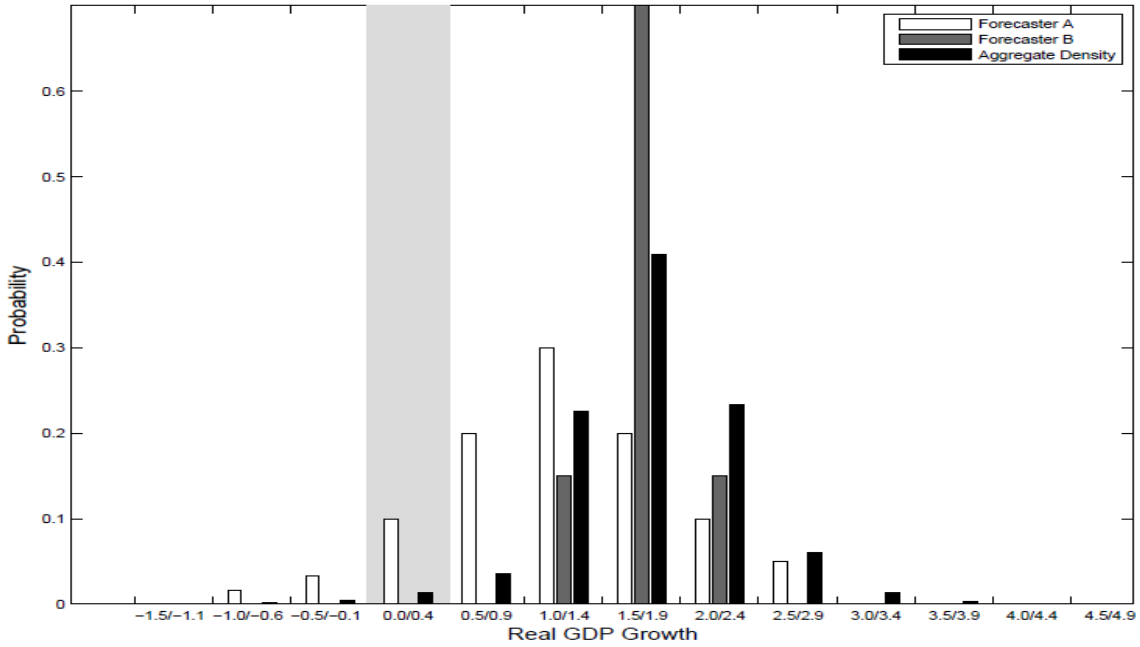
2. EVALUATING PROBABILITY FORECASTS

The traditional approach to testing the information content of density forecasts in macroeconomics has been to use the Probability Integral Transform (PIT) as surveyed in Diebold, Gunther and Tai (1998) and following earlier work by Dawid (1984) and Rosenblatt (1952). However, the latter is particularly problematic in small samples, as is usually the case in macroeconomic applications, given that the null hypothesis of a well behaved density only yields large sample predictions for the statistical properties of the PIT. In addition, for multi-period forecasts, under the null that the predictive density coincides with the true but unobserved density, the distributional features of the PITs are not well defined. As a result it is only possible to test the null of a well behaved density forecast jointly with an assumed model of the process driving the associated multi-period PITs (see, for example, Dowd 2007). As an alternative to the PIT framework for assessing the information content of a given density forecast, we propose examining tests of equal predictive ability between the subjective expert forecasts and a pre-defined benchmark forecast using the framework of Diebold and Mariano (1995) and as extended in Giacomini and White (2006). In particular, we examine a null hypothesis (H_o) of equal predictive ability between the surveyed densities and a benchmark density forecast *conditional* on the information set available to forecasters at the time the forecast was made.[7] To the extent that the surveyed density forecasts have lower overall loss relative to the chosen benchmark, rejection of H_o would represent statistical evidence supporting the marginal information value in the

[7] Giacomini and White (2006) argue that this notion of conditional predictive ability more closely represents the real time problem confronting forecasters in practice.

density forecasts and risk perceptions of macroeconomists. In the remainder of this section, we discuss our design of this test with a view to its subsequent application to the surveyed densities from the ECB SPF.

Figure 1: Individual and aggregate density forecasts for 2008Q3 GDP Growth 1-year ahead (*Annual % Change)*



Note: Forecaster A and B provide examples of two individual density forecasts taken from the survey round conducted in January 2008. The *Aggregate Density* is computed as the equal weighted combination of all individual density forecasts collected in the same survey round. The *shaded region* indicates the range (or survey bin) in which the outcome occurred.

2.1 Evaluation framework

In our evaluation setup, we consider the probability forecasts, $f^{k}_{i,t+\tau}$ produced at time t of individual expert i (i= 1, ..., N) defined over a finite set of outcome ranges (or "bins") indexed from k = 1, . . . K^{MAX} for the forecast target variable of interest in period $t+\tau$, denoted $y_{t+\tau}$.[8] Each $f^{k}_{i,t+\tau}$ represents the probability that expert i assigns to the event that the target variable will fall within the range covered by bin k.[9] Figure 1 gives an illustration of the nature of the data under consideration. The density forecasts in question relate to the outcome for GDP in the third quarter

[8] In the SPF dataset, as discussed in Bowles *et al*. (2010), the forecast horizon (τ) in each survey is set one and two-years ahead of the latest observed outcome and therefore differs across variables due to differing publication lags for the release of official HICP and GDP statistics. For example, the one year ahead GDP forecast refers to the annual growth rate two quarters after the survey quarter, whilst the equivalent HICP forecast refers to the annual inflation rate approximately 11 months after the survey month. For notational convenience, we nonetheless refer to these variable 1 and 2 year ahead rolling horizon forecasts as H=1 and H=2 respectively.

[9] We index $k = 1, \ldots K^{MAX}$ where $K^{MAX} = \max\{K_t\}$ for $0 \leq t \leq T$, where T is the sample size and K_t denotes the actual number of bins used in any given survey (which can change over time). We then conduct the evaluation assuming a zero probability for any range of outcomes not included in a particular survey round.

of 2008 (2008:Q3) which were surveyed in the first quarter of the same year (2008:Q1) when the most recently available data referred to the third quarter of 2007. The Figure portrays two individual density forecasts (denoted A and B, respectively), along with the Aggregate Density forecast from the same survey.[10] The actual outcome relating to these forecasts, as indicated by a shaded region, fell within the range of 0.0-0.4%. The Figure highlights a number of important features of such surveyed forecasts. Forecaster A has spread his predicted probabilities over 8 bins, assigning some positive probability (approx 10%) to the actual outcome that prevailed. Forecaster B, on the other hand demonstrates considerably greater confidence in his predictions, spreading his probability forecasts over only 3 bins. When assessed against the actual outcome as indicated by the shaded region in the chart, in this particular instance, Forecaster B's confidence appears quite ill-judged as he assigns a zero probability to an event that actually occurred and a 70% probability to growth being in a relatively small range (between 1.5% and 1.9%) which turned out to be quite far away from the outcome. The aggregate density forecast, which is based on an equal weighted sum of all probability forecasts from a given survey round, exhibits properties more similar to Forecaster A. It appears overall reasonably symmetric for this particular survey round but has, for example, considerably fatter tails relative to Forecaster B. The aggregate density is also more spread out, i.e. over 10 intervals rather than 8 and 3 intervals for Forecaster A and B, respectively.[11] To evaluate the expert probability forecasts ($f^{k}_{i,t+\tau}$), we assume some function $L(\bullet)$ which can be interpreted in a statistical sense as a measure of overall accuracy and also in an economic sense as a measure of the loss in utility associated with using a given density forecast in a decision theoretic framework. Using $LD_{t+\tau}$ to denote the Loss Differential between the expert density forecast *i* and a candidate benchmark model *j*, the test of equal predictive ability examines the null hypothesis (H_0) that, *conditional* on information set Ω_t , the expectation of the loss differential is zero, i.e.

$$H_0 : \; E[LD_{t+\tau} | \Omega_t] = 0 \tag{2.1}$$

where,

$$LD_{t+\tau} = L\left(f^{k}_{i,t+\tau}\right) - L\left(f^{k}_{j,t+\tau}\right) \tag{2.2}$$

[10] The reported aggregate density is an equal weighted combination of all the individual probability forecasts in a given survey round.

[11] Section 3 of the paper provides more comprehensive information on the sample properties of the individual and aggregate densities.

To test (2.1) on a given forecast in practice, a "test function" (h_t) is used which represents an Ω_t-measurable and q-dimensional vector of information variables that were available at the time the forecasts were made. Under the null of equal predictive ability, (2.1) implies $E[LD_{t+\tau}\, h_t] = 0$. For $h_t = 1.0$, the test reduces to the asymptotic test of equal predictive ability suggested by Diebold and Mariano (1995). Giacomini and White (2006) propose a Wald-type test under which the null has a χ^2 distribution with q degrees of freedom. In the case of multi-period forecasts, it is necessary to base inference on an adjusted variance matrix for $[LD_{t+\tau}\, h_t]$ to account for the resulting serial correlation in the observed loss differentials.[12]

2.2 Scoring the densities

A key element in taking the above framework to the data is the definition of the appropriate loss function or rule, denoted by $L(\bullet)$, in order to "score" the observed accuracy of the density forecasts. Many possible functional forms can be considered, although there may be some prior arguments favouring one form over another.[13] In the economics literature, where continuous densities are the norm, the log predictive score due to Good (1952) has been the dominant scoring rule used for evaluation. It has, in particular, been used to derive performance-based density forecast combinations as in Geweke and Amisano (2011) and Hall and Mitchell (2007). The log score is however not a real valued function and it is therefore less useful for discrete densities of the type depicted in Figure 1.[14] For example, it is undefined if the probability forecast for the actual outcome is zero as was the case with Forecaster B in the preceding illustration. Motivated by these considerations, Boero, Smith and Wallis (2011) have recently emphasised alternative scoring rules such as the quadratic probability score due to Brier (1950) and the Ranked

[12] In the economic context, forecast evaluation systems such as the one proposed here, are complicated by the potential impact of forecasts on policy decisions and thus also on future economic outcomes. In particular, a relatively bad performance of a given forecast in relation to the actual future outcome may not necessarily be an indication of poor forecaster skill but rather it may be consistent with relatively strong forecaster skill which influences policy decisions which in turn have an impact on actual economic outcomes. This ambiguity in the source and interpretation of forecast errors is particularly relevant to longer horizon forecasts where the scope for forecast-based policy changes to impact on actual future realisations is higher. It may also be particularly relevant to density forecasts where policy actions can prevent the materialisation of certain macroeconomic risks if they are credibly identified ex ante in the probability assessments of experts.

[13] One commonly emphasised and desirable feature for such a scoring rule is that it is "proper" in the sense that it would encourage a forecaster to reveal his or her true beliefs or, equivalently, the true density if it were known to him. All of the scoring rules discussed in this paper are proper in this sense. In the case of a strictly proper scoring rule, a forecaster can obtain the unique minimum loss by revealing the (assumed to be known) true density. Gneiting and Raftery (2007) provide a review of various proper scoring rules and their properties.

[14] From an information theory perspective the fact that the log score is undefined for zero probability events that occur is not necessarily a disadvantage. Rather it is consistent with the idea that a rational forecaster should only assign a zero probability to events which are truly impossible. In Section 3, we provide some sample statistics on these "zero probability" events and demonstrate their high incidence in the case of the SPF densities. As highlighted by Boero, Smith and Wallis (2011) this could be overcome by assigning some arbitrarily large value score on such occasions. However, these authors also argue that this is a very unsatisfactory solution since the evaluation and ranking of competing forecasts will be sensitive to such an arbitrarily chosen value.

Probability Score (RPS) due to Epstein (1969). In line with Boero, Smith and Wallis (2011) we favour the use of the RPS given its "sensitivity to distance" as discussed further below. In the case of the discrete probabilities associated with a survey such as the SPF, and using $x^{k}_{t+\tau}$ to denote the binary random variable taking a value of 1 if the period $t+\tau$ outcome occurs in "bin" k and zero otherwise, the sample (of size T) mean values for the RPS is given by (2.2) below.

$$RPS_i = \frac{1}{T-\tau}\sum_{t=1}^{T-\tau}\sum_{k=1}^{K^{MAX}}\left(F^{k}_{i,t+\tau} - X^{k}_{t+\tau}\right)^2 \qquad (2.2)$$

The RPS thus has a quadratic form similar to the quadratic probability score but it is based on the cumulative distribution functions for the candidate densities and the binary outcome variable denoted by upper case $F^{k}_{i,t+\tau}$ and $X^{k}_{t+\tau}$ in (2.2). As a result, the RPS will tend to penalise less severely density forecasts which assign relatively larger probabilities to outcomes that are close to the outcome that actually occurs. It thus gives some relative reward to a density forecast that has a "near miss" while the quadratic probability score will not distinguish between two competing density forecasts in this way.[15] This sensitivity of the RPS to distance is particularly appealing for analysing the impact of density features on overall forecast performance. For example, for a given probability mass assigned to the outcome that occurs, the RPS will reward higher moment features such as skew or thickness of the tails if such features give rise to greater probability mass in regions that are close to where the outcome occurs.

2.3 Benchmark forecasts

Another key ingredient in the proposed test of the predictive content is the definition of the appropriate benchmark forecasts. We propose three benchmarks, graduated in terms of their relative sophistication and differing in terms of the calibration of their means (location) and variances (spread). One reason to consider several benchmarks is to avoid basing inference on a single benchmark which might be inappropriate given the time series properties of the forecast target variable. This would then tend to bias the results in favour of the SPF densities. In order to ensure a "fair" basis of comparison with the expert densities, in making such calibrations we pay particular attention to the real-time character of the benchmarks. In other words, each benchmark

[15] While the concept of distance-sensitive scoring rules is certainly appealing, Gneiting and Raftery (2007) highlight the differing views on the merits of local versus distance–sensitive scoring rules. For example, in contrast to Boero, Smith and Wallis (2011), Bernardo (1979) adopts a perspective which would favour the log score, arguing "when assessing the worthiness of a scientist's final conclusions, only the probability he attaches to a small interval containing the true value should be taken into account".

is constructed in pseudo real time and is thus based on information that would have been available also to an expert forecaster participating in the SPF. To the extent that experts and professional economists can forecast macroeconomic risks, they should tend to outperform all of these suggested benchmarks, given the latter's very mechanical nature. By testing whether or not the SPF density forecasts have any incremental predictive ability compared with such crude real-time statistical models, we thus provide some assessment of the intrinsic "information value" of SPF density forecasts beyond what a simple and largely mechanical tool can provide. Each of the three proposed benchmarks and their calibration is described in more detail below.

Benchmark 1: Uniform forecaster

One natural benchmark against which to assess expert density forecasts is the uniform distribution which assigns an equal probability to all ranges for the outcome variable under consideration. Such a benchmark represents a rather impartial, or even agnostic, perception of future risks. Several practical complications arise in constructing a real time uniform benchmark for surveys such as the SPF. One complication is that the lower and upper tails of the surveyed distributions include open intervals measuring the probabilities of the target variable taking on values above or below a particular threshold. For a practical application, this leads to the need to make some assumption concerning the width of these open intervals. Moreover, the threshold values delineating the open intervals have tended to change over time and, given that each outcome range is of a fixed length (v), the total number of outcome ranges or bins in a given survey round (K_t) has also been modified at various points over time. For example, in early 2009, the number of ranges for the GDP probability forecasts was expanded to include 10 additional outcome "bins" and the lower threshold value was decreased from -1.5% to -6.0%. These modifications in the survey questions reflected actual changes in the economic context which meant that it was natural to modify the structure of the questionnaire. In constructing a Uniform benchmark, we take account of such changes in the range of outcomes under consideration when calibrating the benchmark probabilities. In this way, the benchmark forecasts will approximate an agnostic forecaster who assigns equal probability to each of the outcome ranges that were included in the survey questionnaire at the time it was carried out. The benchmark probabilities are constructed under the simplifying assumption that the open intervals are closed intervals also of length v.[16] Denoting by a_t and b_t the values of the (time-varying) thresholds at the lower and

[16] This assumption follows Zarnowitz and Lambros (1987). Another assumption that has been used in applied studies using such survey data is to equate the open intervals with two equivalent closed intervals. This change would only moderately increase the variance of the uniform benchmark compared with our assumption and we have investigated any impact on our results from varying this assumption and found little sensitivity.

upper ends of the surveyed distribution respectively, the benchmark forecasts are assumed to be uniformly distributed over the interval [a_t - v, b_t + v] and will have probability mass function represented in (2.5) below.

$$f_U(y_{t+\tau}) = \begin{cases} \frac{1}{K_t} & if\ \ a_t - v < y_{t+\tau} < b_t + v \\ 0 \quad otherwise & \end{cases} \tag{2.3}$$

which has a mean ($\mu_{U,t+\tau}$) and variance ($\sigma^2_{U,t+\tau}$) given in (2.6) and (2.7).

$$\mu_{U,t+\tau} = \frac{a_t + b_t}{2} \tag{2.4}$$

$$\sigma^2_{U,t+\tau} = \frac{(b_t - a_t + 2v)^2}{12} \tag{2.5}$$

In (2.3) the total number of bins in a given survey round (K_t) is given a time subscript to stress its variation over time in line with the questionnaire design. (2.3) makes clear the dependence of the uniform benchmark probabilities on the number of bins included in each survey. In particular, the uniform probability assigned to each range will be inversely related to the number of surveyed outcome ranges. As highlighted by (2.5), the variance of the Uniform benchmark is also an increasing function of the distance between the upper and lower thresholds and, hence, with fixed bin width v, it is also increasing with the number of bins used. In line with this, the mean and variance of the benchmark distribution will vary over time in response to these changes in the survey design. This calibration around the actual questionnaire is attractive as it imparts a real time character to the benchmark. The alternative option of calibrating the benchmark probabilities only for a fixed number of bins would seem to be quite arbitrary and would most likely not approximate well an agnostic reply to the questionnaire.[17]

Benchmark 2: Gaussian random walk

A second simple benchmark density forecast is calibrated on the basis of the Gaussian distribution and is based on the premise that the forecast target variable follows in levels a

[17] Although the uniform benchmark probabilities are computed based on a time-varying K_t, the probability scores are computed by summing over a constant number of bins which is taken to be the largest number of surveyed bins used during the sample period, i.e. K^{MAX} = max $\{K_t\}$ for $0 \leq t \leq T$, where T is the sample size. For bins that are not explicitly surveyed, in calculating the scores a zero probability forecast is assumed when computing the benchmark scores. This avoids distorting the performance measures simply through the use of a different number of ranges for the surveyed and benchmark probability forecasts.

random walk with drift. Although less agnostic than the uniform, the choice of a Gaussian distribution reflects its widespread use in macroeconomics as a basic statistical representation for many time series such as the price level or the level of GDP. In addition, the probability forecasts from the random walk benchmark exhibit properties of symmetry and they imply an absence of large tail risks. As both features are likely to be important aspects of the true underlying uncertainty inherent in the data generating process, especially given the inclusion of the recent period of macroeconomic instability in our evaluation sample, the comparison with a Gaussian random walk offers experts in the SPF ample scope to demonstrate any subjective skill in their risk assessments. The proposed second benchmark therefore is derived from a normal distribution $N(\mu_{rw,t+\tau}, \sigma^2_{rw,t+\tau})$ with mean $\mu_{rw,t+\tau}$ and variance $\sigma^2_{rw,t+\tau}$. When estimated on quarterly data expressed in log levels, the drift parameter (α_t) of the random walk approximates the historical sample mean quarterly growth rate of the time series at the time the survey was conducted.[18] The mean of the Gaussian benchmark is calibrated as an annual growth rate consistent with this drift parameter and allowing for publication lags in order to preserve the real time nature of the comparison. A corresponding real time estimate of the variance of the predictive density can then be calibrated on the basis of the *ex post* point forecast errors implied by this benchmark's mean forecast. The mean and variance of the predictive density for the Gaussian random walk benchmark will then be given by (2.6) and (2.7) below.

$$\mu_{rw,t+\tau} = 4\alpha_t \tag{2.6}$$

$$\sigma^2_{rw,t+\tau} = (t-\tau+1)^{-1} \sum_{t=0}^{t-\tau} \left(y_{t+\tau} - \mu_{rw,t+\tau} \right)^2 \tag{2.7}$$

As with the uniform distribution, the variance of the above Gaussian benchmark will adjust mechanically in response to changes in the economic environment.[19] For example, the large forecast errors associated with the 2008-2009 recession provoke a significant increase in the variance implied by (2.7). In order to derive the survey replies consistent with the Gaussian random walk, we simulate 10,000 draws from a Gaussian distribution with mean and variance

[18] The random walk with drift is estimated from the regression $\Delta y_t = \alpha_t + u_t$ where Δy_t refers to the first difference of the natural log of the seasonally adjusted level of the forecast variable (y_t) and u_t represents the equations residuals.

[19] Clark (2011) highlights the challenges to density forecasting that arise from sharp changes in macroeconomic volatility, finding that adding stochastic volatility to a Bayesian Vector AutoRegression (BVAR) model for US GDP materially improves the real time accuracy of point and density forecasts.

given by (2.6) and (2.7) above and estimate the corresponding $f^{k}_{rw,t+\tau}$ as the proportion of draws falling within a given outcome range.[20]

Benchmark 3: Naive market forecast

In contrast to Benchmark 1 and 2 which are based on crude distributional assumptions, a final proposed benchmark uses information from the SPF itself. In completing the survey, one simple strategy that respondents could adopt is to use the most recently observed aggregate density forecast, i.e. as taken from the results of the previous survey round which were publicly available.[21] We therefore propose a "naïve market" benchmark forecast that is motivated by the idea that the most recently observed aggregate probability forecast, which was publicly available and therefore accessible by all forecasters participating in the survey, represents another low lying threshold against which to assess individual expert opinion. In particular, to the extent that professional forecasters can efficiently exploit any news associated with the information flow on macroeconomic developments between survey rounds, they should be able to generate better density forecast performance compared with this benchmark. The naïve market forecast is thus given by the equal weighted aggregate of the N individual replies to the most recently observed survey round and has probability mass function given by (2.8) below.

$$f_{NM}(y_{t+\tau}) = N^{-1}\sum_{i=1}^{N} f_i(y_{t+\tau-1}) \tag{2.8}$$

An important feature of (2.8) is that it does not involve any strong prior assumption on the functional form of the benchmark distribution. Indeed, as a finite mixture distribution, the linear opinion pool represented by (2.8) combines the information in the shape of individual density forecasts with information on the disagreement among individuals on the location of those densities. The mean and variance of this benchmark are given by (2.9) and (2.10) below.

$$\mu_{NM,t+\tau} = N^{-1}\sum_{i=1}^{N} \mu_{i,t+\tau-1} \tag{2.9}$$

[20] In contrast to the calibration of the uniform distribution, we do not constrain the Gaussian random walk benchmark probabilities to those ranges included in each questionnaire. The reason for this is that the latter approach would result in a non-monotonic and often highly skewed benchmark density given that the benchmark mean and variance would imply very large probabilities in the outcome ranges at the extremities of the ranges that were actually surveyed. The Gaussian benchmark therefore has non-zero probabilities also in those outcome ranges that were not explicitly included in the survey questionnaire.

[21] In particular, the results of each quarterly SPF round are published on the ECB website and in the ECB Monthly Bulletin approximately two to three weeks after the completion of each survey round.

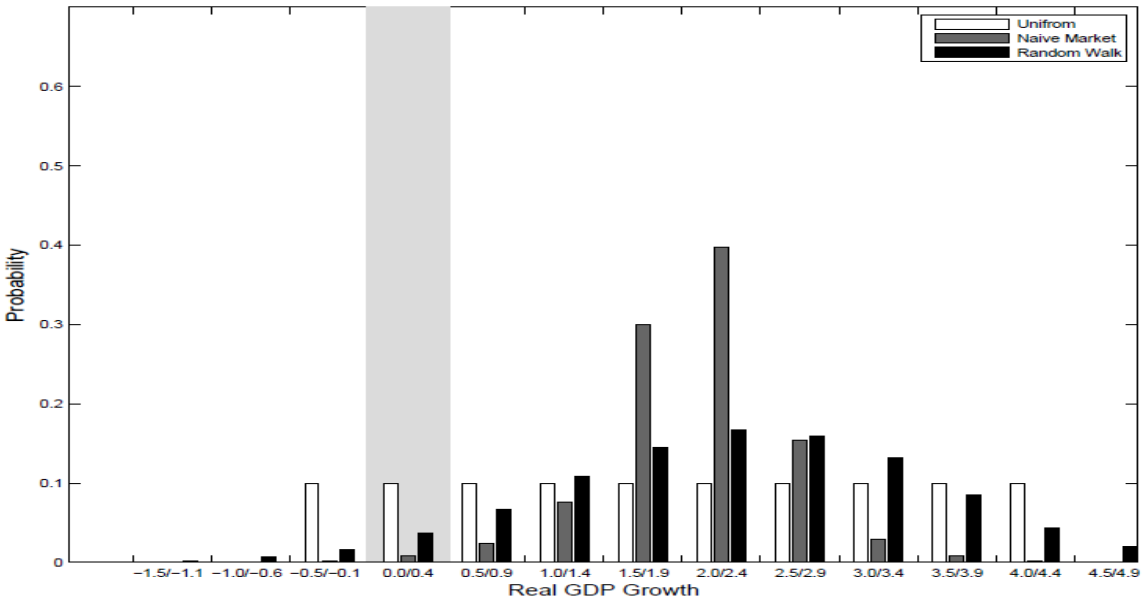
$$\sigma^2_{NM,t+\tau} = N^{-1}\sum_{i=1}^{N}\sigma^2_{i,t+\tau-1} + N^{-1}\sum_{i=1}^{N}\left(\mu_{i,t+\tau-1} - \mu_{NM,t+\tau-1}\right)^2 \tag{2.10}$$

According to (2.9), the mean of the naïve market forecast is simply the equal weighted sum of the individual means in the most recently observed survey round. However, as a combined density forecast, its variance will tend to be higher than the average of the individual variances given by the first term on the right hand side of (2.10) whenever there is disagreement across individual means as captured by the second term on the right hand side of (2.10). Such disagreement could, for example, reflect divergent expert views concerning the appropriate model underpinning the economy's dynamics.

To get some basic sense into the properties of the different benchmark forecasts, Figure 2 provides an illustration of the predictive distributions for each of the three benchmarks discussed above.[22] Figure 2 - like Figure 1 - relates to the specific case of the probability forecasts for GDP growth in 2008:Q3. As reflected in the chart, each of the densities had a relatively similar mean for this specific quarter. However, they differ substantively in terms of other features. The uniform, also in general, has a substantially higher variance, a feature which proved helpful for the quarter in question as it was consistent with a relatively high probability (10%) assigned to the actual outcome as indicated by the shaded region in the chart. Comparing the random walk with the naïve market based forecast also highlights some notable differences in shape. In particular, the survey-based benchmark tends to have a considerably lower variance relative to the random walk calibrated on historical forecast errors. This is reflected in a considerably higher probability assigned to the outcome ranges toward the mid of the distribution, and correspondingly, a relatively low probability assigned to the actual outcome which was more toward the extremes of the distribution. Finally, in contrast to the uniform and random walk benchmarks which are fully symmetric, a notable feature of the naive benchmark is its clear skew which was negative for this specific quarter. These broad differences in the benchmark density forecasts for this specific quarter also generalise to the full sample (see the further discussion in Section 3.2 below).

[22] In section 3, we provide some further summary statistics over the full sample period and background information on the key properties of the benchmark forecasts, also in comparison with the corresponding individual and aggregate SPF density forecasts.

Figure 2: Sample benchmark density forecasts for 2008Q3 GDP Growth 1-year ahead (*Annual % Change*)



Note: The *Uniform* benchmark gives equal probably mass is attached to all outcome ranges included in a given survey round. The *Random Walk* benchmark is a Gaussian density forecast calibrated under the assumption that the level of the forecast target variable follows a random walk with drift and with predictive variance calibrated on the basis of observed past forecast errors. The *Naïve Market* benchmark is the aggregate density forecast from the previously published survey round. The *shaded region* indicates the range (or survey bin) in which the outcome occurred.

3. THE SPF DENSITY FORECASTS

The SPF dataset used in the study is quarterly and refers to the surveys conducted over the sample period 1999Q1-2011Q1. The sample thus includes a number of influential observations linked to the financial crisis which was associated with the "great recession" of 2008-2009. The complete micro dataset for the euro area SPF conducted since 1999Q1 can be directly downloaded at http://www.ecb.europa.eu/stats/prices/indic/forecast/html/index.en.html. Data for euro area GDP growth and inflation outcomes, which are used for the density evaluation and for the calibration of the benchmark densities, are obtained from Eurostat, the official statistical institute of the EU. GDP growth is measured using chain linked volumes based on previous year prices, while inflation is measured by the Harmonised Index of Consumer Prices for the euro area. In the remainder of this section, we provide some additional background information on the survey data on which our study draws. In particular, we focus on explaining how we have handled gaps in the dataset relating to the irregular participation of respondents in the surveys as well providing some descriptive information and summary statistics in order to highlight key features of the surveyed distributions.

3.1 Irregular participation and non-responses

One practical challenge that emerges in empirical analysis with such survey data is the irregular participation of survey participants. In addition, even respondents who reply on a regular basis may on occasion fail to submit a reply. Such irregular participation and non-responses result in an unbalanced panel which poses challenges to econometric estimation and hypothesis testing and, thus, could impact significantly on forecast evaluation. For example, forecasters which did not submit replies during the recent period of macroeconomic volatility could be favoured if the performance statistics were based on sample averages without any adjustment for these gaps in

Table 1: Proportion of non-replies in filtered and non-filtered data sets
(% of total number of surveys, 1999Q1 – 2011Q1)

	GDP Growth				Inflation			
	(H=1)		(H=2)		(H=1)		(H=2)	
Min	2%	(2%)	2%	(2%)	2%	(2%)	2%	(2%)
10%	8%	(6%)	10%	(6%)	10%	(6%)	10%	(6%)
25%	18%	(8%)	24%	(8%)	18%	(8%)	22%	(8%)
50%	42%	(14%)	50%	(12%)	40%	(12%)	52%	(12%)
75%	76%	(20%)	80%	(24%)	74%	(20%)	78%	(22%)
90%	90%	(26%)	94%	(28%)	88%	(26%)	90%	(30%)
Max	98%	(36%)	98%	(54%)	98%	(38%)	98%	(38%)
N	96	(27)	95	(27)	96	(28)	96	(28)

Note: The percentage of non-responses in the filtered data is given in ( ). N denotes the number of individual density forecasts in the raw (and filtered) data sets respectively.

the data. Some simple solutions to this entry and exit problem have been highlighted in Capistrán and Timmerman (2009) for the case of point forecast, while Boero, Smith and Wallis (2011) have recently proposed a correction for the computation of probability scores in the presence of such irregularities. In the case of the raw SPF data set, the problem of irregular participation is particularly acute for density forecasts, with – depending on the particular forecast variable or horizon - the median number of non responses in the region of 40% to 50% of the total number of surveys conduced over the sample period.

Our approach to this issue follows Genre *et al.* (2010) in constructing a fully balanced panel from the raw SPF dataset which allows us compare each individual SPF forecaster with the benchmark

forecasts over an identical sample period. Given the large number of missing observations in the raw dataset, any attempt to estimate all of them would risk introducing an excessive element of estimation error into the subsequent evaluation results. In preparing this fully balanced panel, we therefore proceed in two steps which are aimed at trading-off the possible introduction of some sampling error whilst limiting the impact of estimation error in balancing the panel. In a first step, we sample from the raw SPF panel and employ a decision rule to filter out highly irregular survey participants. We define these irregular participants to be those with more than four consecutive non replies. Table 1 summarises the results of this filtering. It reduces considerably the proportion of missing observations across the panel and therefore, as argued above, should contribute to reducing the impact of estimation error when comparing performance across individual forecasters. An unavoidable downside of such an approach, however, is a substantial reduction in the number of individual densities considered. In the case of GDP forecasts 1 year ahead, for example, the sample size is reduced from over 90 to 27 individual respondents. Nonetheless, we would argue that such a sample remains quite representative of the population of euro area macroeconomic experts and includes forecasters from a broad cross section of euro area and some EU countries, including the UK. For the filtered dataset, in a second step, we balance the remaining panel of density forecasts by estimating the location (mean) of the missing densities as done in Genre *et al*. (2010) using a panel regression which estimates the degree of persistence in the deviation of individual point forecasts ($\mu_{i,t+\tau}$) from the equal weighted average point forecast $\overline{\mu}_{t+\tau}$.

$$\mu_{i,t+\tau} - \overline{\mu}_{t+\tau} = \beta(\mu_{i,t+\tau-1} - \overline{\mu}_{t+\tau-1}) + \varepsilon_{i,t+\tau} \tag{3.1}$$

Using an estimate of β, this expression can be used to update point forecasts for which no survey reply is available. For a non-responding panel member, this updated point forecast can in turn be used to estimate the "bin" containing the centre of the forecaster's updated distribution. In particular, a missing mean forecast for period *t* can be computed conditional on the equal weighted forecasts in period *t* and *t*-1 and the previously reported individual point forecast for period *t*-1. Conditional on this estimate of the missing mean forecasts, the individual densities are then centred on the outcome range or bin containing this point forecast and the remaining probabilities are filled in using the probabilities from the density forecast reported in the most recent period. For example, consider forecaster *i* who assigned the probabilities $f^{3}_{i,t+\tau}$ = 20%, $f^{4}_{i,t+\tau}$ = 60%, $f^{5}_{i,t+\tau}$ = 20% to bins 3, 4 and 5 respectively (and zero probability to all other bins) but who does not respond to the survey in period *t*+1. The missing *t*+1 density would then be

centred on the bin containing the updated mean using equation (3.1) but using the period t probabilities. For example, if the updated mean forecast was an element of bin 3, then the t+1 density forecast would be given by $f^{2}_{i,t+\tau+1}$ = 20% , $f^{3}_{i,t+\tau+1}$ = 60% and $f^{4}_{i,t+\tau+1}$ = 20%. As can be seen from Table 1, in the filtered data set for the more reliable participants only 1 missing density forecast needs to be estimated in this way (or approximately 2% of the surveys conducted over our sample). Depending on the forecast variable and horizon, for the forecaster with the median number of non-responses, between 12 and 14% of the survey replies need to be estimated in this way. However, for the least reliable "serial non-respondents", the proportion of missing densities that needs to be estimated in the filtered dataset is considerably higher though it rarely exceeds 40% of the total number of surveys that have been conducted. Nonetheless, a detailed examination of the micro data suggests we can be reasonably confident on the accuracy of this method to balance out the dataset. In particular, such an examination suggests that forecasters generally do not change rapidly the shape of their density forecasts from one period to the next. Rather, looking at the professional forecasters' responses, the evidence suggests there tends to be considerable persistence in the shape of the individual surveyed density forecasts over time.

3.2 Features of the surveyed densities

Tables 2 and 3 provide more complete summary information on the panel of densities for both GDP growth and consumer price inflation. In particular, the tables report for the one and two year ahead (H=1 and H=2 respectively) GDP growth and inflation density forecasts, several key summary statistics: 1) the point forecast as estimated by the mean (μ) of the predictive density, 2) the Mean Error (ME) and Mean Absolute Error (MAE) of the equivalent point forecast, (3) the estimated variance (σ^2) of the density forecast (4) the *absolute* skew (γ) of the density forecast,[23] (5) and the degree of excess of kurtosis (ζ) in the surveyed density forecast. We follow the approach suggested in Lahiri, Teigland and Zaporowski (1988) in order to calculate these summary statistics. All summary statistics are reported in terms of their sample means taken over the period 1999Q1-2011Q1. Each distributional characteristic is reported for the cross section of individual replies as summarised by the Median (50th percentile), the Max/Min, the 10th/90th and 25th/75th percentiles of the distribution. For comparison, in each table equivalent summary statistics are also reported for the three benchmark density forecasts described in Section 2.

[23] We focus on the average absolute skew in order to provide a sense of the overall degree of asymmetry in the expert densities, whether positive or negative. Simply taking the average actual skew could imply relatively symmetric distributions on average given that periods of positive skew would tend to be offset by periods of negative skew.

Surveying the summary statistics in Tables 2 and 3 highlights positive mean errors for GDP and negative mean errors for inflation, indicating a tendency for the individual density forecasts to over predict GDP growth and under predict inflation during the sample period. In terms of overall point forecast accuracy of the GDP growth forecasts (Table 2) as measured by the MAE, all forecasters outperform the most accurate of the three benchmarks (which is the uniform). At the 2-year horizon, the equivalent proportion is notably lower with only a few individuals outperforming the most accurate benchmark. In the case of inflation (Table 3), the best performing benchmark forecast according to the MAE is the random walk for H=1 and the Naïve Market forecast for H=2. The relative performance of the surveyed expert forecast is somewhat less impressive for inflation compared with growth. Nonetheless for both horizons, a significant

Table 2: Sample statistics: SPF and benchmark density forecasts for GDP Growth
(*Sample Averages, 1999Q1-2011Q1***)**

	μ	ME	MAE	σ^2	γ	ζ
			H=1			
Min	1.26	-0.10	0.99	0.09	0.07	-0.81
10%	1.41	0.05	1.08	0.09	0.10	-0.75
25%	1.62	0.08	1.13	0.13	0.16	-0.62
50%	1.72	0.19	1.21	0.19	0.17	-0.37
75%	1.82	0.27	1.25	0.29	0.25	-0.22
90%	1.90	0.34	1.28	0.38	0.37	0.01
Max	2.00	0.62	1.47	0.79	0.59	0.64
Aggregate SPF	1.68	0.20	1.15	0.42	0.21	1.33
Uniform	1.73	0.24	1.40	2.74	0.00	-1.22
Random Walk	2.27	0.83	1.51	1.89	0.05	-0.10
Naïve Market	1.69	0.25	1.42	0.41	0.22	1.36
			H=2			
Min	1.68	0.40	1.32	0.11	0.06	-0.95
10%	1.78	0.57	1.39	0.12	0.11	-0.78
25%	1.97	0.61	1.53	0.17	0.15	-0.58
50%	2.06	0.87	1.59	0.23	0.20	-0.41
75%	2.10	0.95	1.67	0.37	0.25	-0.29
90%	2.14	1.06	1.72	0.51	0.27	-0.23
Max	2.32	1.14	2.07	1.07	0.47	0.85
Aggregate SPF	2.05	0.86	1.59	0.45	0.29	1.40
Uniform	1.73	0.56	1.53	2.74	0.00	-1.22
Random Walk	2.29	0.98	1.55	1.83	0.07	0.18
Naïve Market	2.05	0.94	1.60	0.45	0.29	1.40

Note: μ denotes the average mean of the predictive density, ME denotes the average error, where the error is defined as the mean forecast (μ) minus the outcome, MAE denotes the mean absolute error, σ^2 denotes the average variance of the density forecast, γ denotes the average *absolute* skew, ζ denotes the average excess kurtosis. All variables are computed as sample averages taken over the period 1999Q1-2011Q1.

proportion of individual forecasters are able to outperform the best performing benchmarks. Turning to the other sample statistics highlighted in Tables 2 and 3, a very clear feature of the surveyed replies is the relatively low variance of the expert density forecasts, in particular relative to the benchmark densities. For example, in the case of GDP (H=1), even the individual with the highest spread in his predictive distribution still has a variance that is only half that of the random walk benchmark calibrated on historical forecast errors. This feature holds across both variables and both horizons. Such low variances in the expert forecasts are suggestive of possible "over-confidence", "local thinking" or "neglected risks", whereby certain states of the world do not

Table 3: Sample statistics: SPF and benchmark density forecasts for Inflation
(*Sample Averages, 1999Q1-2011Q1***)**

	μ	ME	MAE	σ^2	γ	ζ
			H=1			
Min	1.58	-0.59	0.49	0.08	0.05	-0.91
10%	1.63	-0.42	0.51	0.09	0.13	-0.82
25%	1.70	-0.34	0.57	0.13	0.15	-0.56
50%	1.74	-0.30	0.65	0.20	0.18	-0.34
75%	1.80	-0.25	0.69	0.25	0.25	-0.15
90%	1.90	-0.21	0.72	0.34	0.33	0.07
Max	2.01	-0.01	0.78	0.59	0.73	1.54
Aggregate SPF	1.75	-0.30	0.63	0.30	0.10	0.87
Uniform	1.66	-0.39	0.68	1.98	0.01	-1.21
Random Walk	1.80	-0.24	0.63	0.59	0.02	0.02
Naïve Market	1.75	-0.31	0.68	0.30	0.10	0.87
			H=2			
Min	1.51	-0.43	0.48	0.09	0.05	-1.10
10%	1.71	-0.39	0.57	0.10	0.10	-0.66
25%	1.74	-0.31	0.61	0.13	0.13	-0.59
50%	1.79	-0.24	0.67	0.22	0.18	-0.40
75%	1.85	-0.19	0.70	0.32	0.20	-0.26
90%	1.94	-0.15	0.77	0.47	0.29	-0.07
Max	2.09	-0.08	0.78	0.74	0.53	1.09
Aggregate SPF	1.79	-0.25	0.66	0.35	0.10	1.02
Uniform	1.66	-0.36	0.69	1.98	0.01	-1.21
Random Walk	1.82	-0.24	0.65	0.63	0.02	0.36
Naïve Market	1.79	-0.24	0.65	0.35	0.10	1.02

Note: μ denotes the average mean of the predictive density, ME denotes the average error, where the error is defined as the mean forecast (μ) minus the outcome, MAE denotes the mean absolute error, σ^2 denotes the average variance of the density forecast, γ denotes the average *absolute* skew, ζ denotes the average excess kurtosis. All variables are computed as sample averages taken over the period 1999Q1-2011Q1.

"come to mind" when respondents are formulating their replies to the questionnaire.[24] This is very much in line with experimental evidence from behavioural economics and psychology, as - for example - documented in Kahneman and Tversky (2000) and also as documented in earlier studies such as Giordani and Soderlind (2003) who find that the interval forecasts of professional economists for inflation are too narrow.[25] In line with this, it is also notable that the variances of the surveyed densities are only marginally increasing in the forecast horizon, suggesting that macroeconomic experts are almost as confident about their two year ahead predictions as they are about their one year ahead predictions. The individual surveyed densities also generally exhibit negative excess kurtosis, i.e. a lower probability mass in the tails of the distribution relative to a Gaussian distribution with the same variance. For both GDP and inflation, a few individual forecasters do attribute higher tail probabilities on average as indicated by positive excess kurtosis but this is more the exception. From the tables, one can also see the thickening of the probability tails that emerges through taking the linear opinion pool as reflected in the positive excess kurtosis for the Aggregate SPF density. Finally, in contrast to the random walk and uniform benchmarks which are symmetric by construction, the expert densities often exhibit positive absolute skew over the sample period.[26] Given that our sample includes the large macroeconomic volatility associated with the recent financial crisis, it is of interest to consider in the empirical evaluation how these higher moment features of the density forecasts impact on density forecast performance.

The above indication of possible overconfidence in the surveyed expert forecasts is very much in line with the relatively large share of occasions on which respondents assign a zero probability to an event that subsequently occurs. Table 4 provides some summary information on the proportion of times forecasters "completely miss" the outcome in this sense. For each variable and horizon, the table shows this proportion for the cross section of forecasters together with the equivalent proportions from the aggregate and benchmark density forecasts. The incidence of such "neglected risks", as reflected in percentage of rounds where a zero probability was reported for

[24] This result is similar to Giordani and Söderlind (2006) who report evidence of overconfidence in macroeconomic density forecasts for the US SPF. Gennaioli and Shleifer (2010) provide a rigorous formalisation of such behaviour and model the neglect of certain states of the world using the idea of "local thinking" whereby not all contingencies are represented in the decision maker's thought processes. Building on these insights, Gennaioli, Shleifer and Vishny (2010) develop a model which explores the role of neglected risks as an explanatory factor in contributing to the recent financial crisis.

[25] Mitchell and Wallis (2011) conjecture that the recourse to expert judgement as opposed to econometric models is a possible source of this over confidence.

[26] The summary statistics reveal some small skew and kurtosis for the random walk benchmark. This reflects the fact that we have used a discrete approximation over a fixed number of bins (in line with the survey replies) to derive the associated summary statistics.

an outcome that subsequently occurred is remarkably high throughout the SPF panel. Such neglected risks tend to be particularly significant in the case of GDP growth with the median proportion of zero probability events that subsequently materialised standing at 47% and 53% for the 1 and 2-year horizons respectively. The equivalent proportions for consumer price inflation, at 20% and 24%, while clearly lower are nonetheless indicative of important neglected risks by survey participants. These data also point to an increase in such local thinking as the horizon increases, in line with the evidence in Tables 2 and 3 that the predictive variances are only marginally increasing in the forecast horizon. In the case of the benchmark density forecasts, the incidence of neglected risks is considerably lower.

Table 4: Proportion of surveys where zero probability is assigned to an outcome that occurs
(% of total number of surveys 1999Q1-2011Q1)

	GDP growth		Inflation	
	(H=1)	**(H=2)**	**(H=1)**	**(H=2)**
Best	17%	12%	4%	7%
10%	23%	23%	9%	12%
25%	40%	37%	15%	14%
50%	47%	53%	20%	24%
75%	55%	60%	26%	31%
90%	66%	67%	33%	33%
Worst	68%	70%	37%	36%
Aggregate	9%	12%	0%	0%
Random Walk	9%	7%	0%	0%
Uniform	11%	12%	0%	0%
Naïve Market	13%	12%	0%	0%

Note: The table reports the number of times a given forecaster assigns a probability of zero to an event which subsequently materialises expressed as a share of the total number of surveys over the sample.

4. EVALUATION RESULTS

The summary statistics presented in Section 3 suggest that, at the individual level, professional macroeconomists demonstrate a considerably high degree of confidence in the macroeconomic predictions as reflected in relatively small estimated sample variances from their predictive densities and a high proportion of times when they assign a zero probability to an event that subsequently materialises. This excess confidence, which is particularly evident in GDP forecasts

and increases with the forecast horizon, is also reflected in relatively low probability values being assigned to more extreme outcomes, e.g. relative to a Gaussian distribution. An open question remains, however, as to whether such features of the surveyed density forecasts are reflected in a relatively good or a relatively poor overall density forecast performance relative to the benchmarks. In this Section, we therefore turn to a more robust empirical evaluation of the densities using the framework described in Section 2.

Tables 5 and 6 report the quantiles of the distribution of individual ranked probability scores for the GDP and inflation density forecasts at both the one and two year horizons. The scores are reported for the cross section of individual forecasters in relative terms for each of the three benchmarks, i.e. divided by the equivalent benchmark score, with a relative score less than unity indicating an improvement in overall forecast performance compared with the benchmark. In addition, the tables report P-values for each relative score providing an estimate of the conditional and unconditional likelihood that the subjective density forecasts and the benchmark forecasts have equal predictive ability. A number of important insights emerge from this evaluation of individual density performance. A first clear impression is of considerable heterogeneity in density forecast performance at the individual level for the case of both GDP and inflation densities. Depending on the forecasting horizon, a differential of as much as 40% is observed when comparing the relative score of the best and worst performing experts. In line with this, some experts exhibit a density performance that is substantially (and often statistically) worse than several of the crude benchmarks while others demonstrate quantitative improvements in relative terms. The evaluation results in these tables, also generally point to the benefits from combining individual densities for both variables and at both forecast horizons. For example, for all variables and all horizons, the Aggregate SPF combination outperforms a large majority of individual forecasters. In particular, for all variables and all horizons, the Aggregate SPF combination is at least as good as the top 25% of individual forecasters and it is often in the top 10%.[27]

Turning in more detail to the evaluation of the densities for GDP growth in Table 5, the ability of some experts to improve on the benchmarks is strongest for forecasting GDP growth at shorter horizons of one year where 75% of all experts perform better than the best performing of the

[27] However, the best performing density is never the Aggregate SPF combination for any variable or any horizon. This highlights a case for future research to consider alternative approaches to combining the expert densities in the SPF.

Table 5: Relative probability scores and tests of equal predictive ability: GDP Growth Densities

	Uniform			Random Walk			Naïve Market		
	Relative Score	**P-value [h_t=1.0, y_t]**	**P-value [h_t=1.0]**	**Relative Score**	**P-value [h_t=1.0, y_t]**	**P-value [h_t=1.0]**	**Relative Score**	**P-value [h_t=1.0, y_t]**	**P-value [h_t=1.0]**
					Horizon (H) = 1				
Best	0.82	0.46	0.12	0.75	0.45	0.11	0.76	**0.02**	**0.01**
10%	0.87	0.70	0.26	0.79	0.38	0.22	0.80	0.30	**0.07**
25%	0.90	0.63	0.33	0.81	0.48	0.15	0.83	0.31	**0.06**
50%	0.94	0.38	0.66	0.85	0.25	0.42	0.86	0.70	0.23
75%	0.99	0.59	0.91	0.90	0.27	0.44	0.91	0.59	0.17
90%	1.07	0.04	0.51	0.97	**0.06**	0.83	0.98	0.15	0.86
Worst	1.11	0.12	0.06	1.01	0.12	0.94	1.02	0.27	0.62
Aggregate	0.86	0.57	0.17	0.78	0.49	0.15	0.79	**0.00**	**0.01**
					Horizon (H) = 2				
Best	1.05	0.13	0.53	0.96	0.48	0.68	0.92	**0.07**	0.18
10%	1.08	0.06	0.26	1.00	0.22	0.95	0.96	**0.08**	0.17
25%	1.14	0.04	0.16	1.05	0.48	0.17	1.00	0.42	0.78
50%	1.21	0.09	0.02	1.12	0.32	0.14	1.07	0.58	0.18
75%	1.25	0.00	0.01	1.15	0.04	0.00	1.10	0.03	0.01
90%	1.27	0.02	0.01	1.17	0.00	0.00	1.12	0.00	0.00
Worst	1.30	0.02	0.01	1.20	0.00	0.00	1.15	0.01	0.00
Aggregate	1.13	0.05	0.16	1.04	0.65	0.25	1.00	0.80	0.65

Note: Relative score gives the average individual score divided by the average benchmark score. In bold P-values corresponding to individuals with conditional predictive ability higher than the benchmark and with a significance level lower than 10%, y_t indicates the conditioning variables in the test of conditional predictive ability and includes the time series for inflation and GDP growth that were available at the time the survey was carried out.

Table 6: Relative probability scores and tests of equal predictive ability: Inflation Densities

	Uniform			Random Walk			Naïve Market		
	Relative Score	**P-value [h_t=1.0, y_t]**	**P-value [h_t=1.0]**	**Relative Score**	**P-value [h_t=1.0, y_t]**	**P-value [h_t=1.0]**	**Relative Score**	**P-value [h_t=1.0, y_t]**	**P-value [h_t=1.0]**
					Horizon (H) = 1				
Best	0.77	**0.00**	**0.08**	0.93	0.64	0.55	0.84	**0.00**	**0.01**
10%	0.79	**0.02**	**0.02**	0.95	0.96	0.61	0.88	0.37	0.15
25%	0.85	**0.06**	**0.03**	1.02	0.98	0.80	0.94	0.59	0.52
50%	0.93	0.36	0.36	1.12	0.51	0.18	1.03	0.11	0.63
75%	1.00	0.47	0.99	1.20	0.00	0.00	1.08	0.03	0.27
90%	1.02	0.83	0.89	1.23	0.49	0.22	1.13	0.72	0.33
Worst	1.08	0.24	0.43	1.30	0.00	0.00	1.19	0.01	0.04
Aggregate	0.85	**0.01**	**0.06**	1.02	0.96	0.77	0.93	0.19	**0.10**
					Horizon (H) = 2				
Best	0.81	0.24	0.15	0.92	0.61	0.42	0.91	0.62	0.41
10%	0.87	0.28	0.56	0.99	0.34	0.94	0.98	0.28	0.56
25%	0.90	**0.01**	0.46	1.03	0.10	0.83	0.99	0.40	0.92
50%	0.97	0.13	0.80	1.10	0.44	0.12	1.08	0.10	0.11
75%	1.02	0.01	0.89	1.16	0.37	0.09	1.13	0.00	0.00
90%	1.08	0.43	0.66	1.22	0.56	0.17	1.20	0.03	0.12
Worst	1.13	0.16	0.32	1.29	0.00	0.00	1.27	0.00	0.00
Aggregate	0.90	**0.08**	0.43	1.02	0.63	0.79	1.00	0.35	0.91

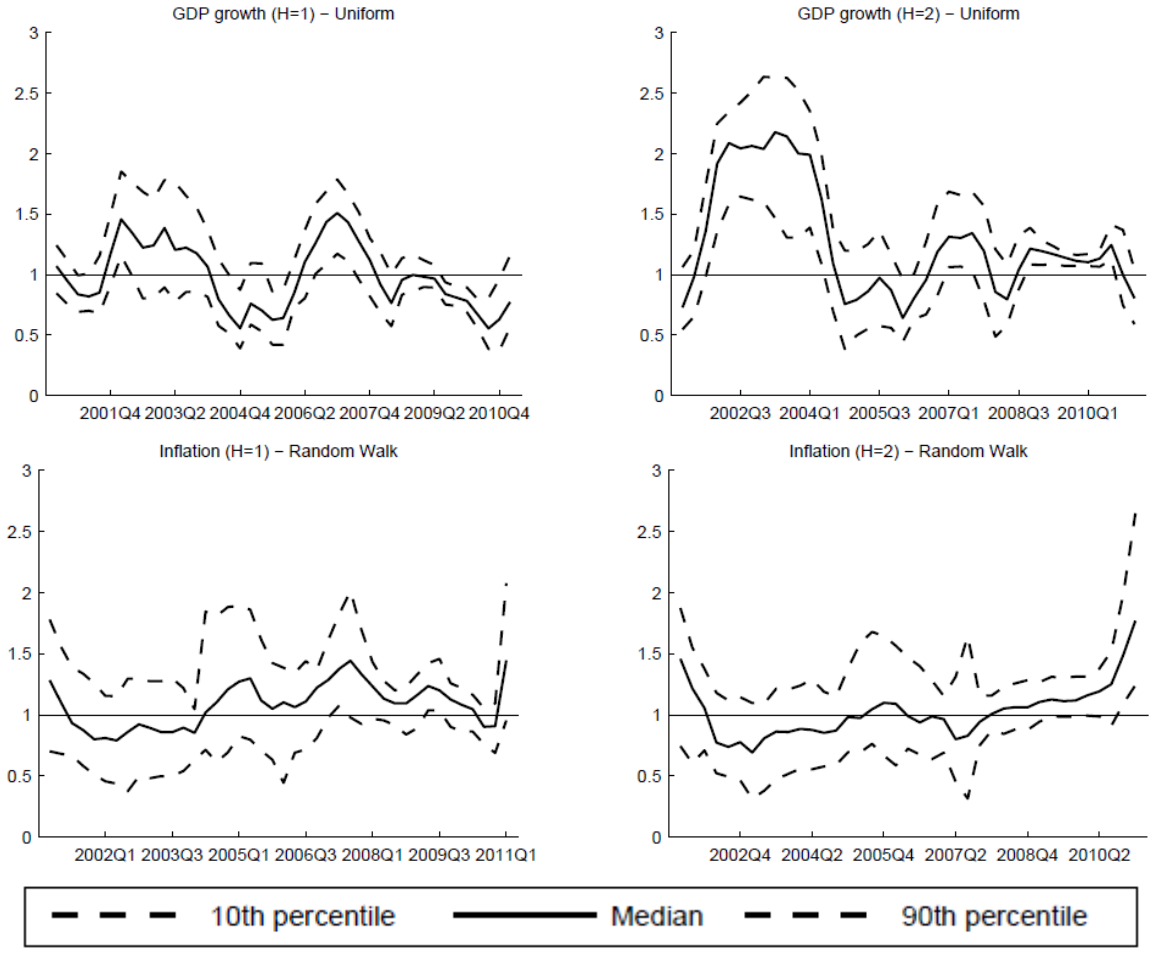
Note: Relative score gives the average individual score divided by the average benchmark score. In bold P-values corresponding to individuals with conditional predictive ability higher than the benchmark and with a significance level lower than 10%, y_t indicates the conditioning variables in the test of conditional predictive ability and includes the time series for inflation and GDP growth that were available at the time the survey was carried out.

mechanical benchmarks. However, according to the tests of predictive ability, rarely are these quantitative improvements significant in a statistical sense. Moreover, at longer forecasting horizons of two years, the quantitative gains for the best performing experts for GDP growth is lost. Indeed, for GDP growth, an agnostic uniform benchmark performs better than any of the experts in the panel. In addition, for the 2 year horizon, the majority of forecasters (at least 75% of all experts) fail to perform better than either the random walk or the naïve market forecast at this longer horizon, while, for those that do, the gains are quantitatively very small. Overall, therefore, this points to a relatively sharp deterioration in the information content of even the best performing expert densities as the forecast horizon is extended from 1 to 2 years.

In the case of the inflation densities in Table 6, the performance of the expert risk assessments is somewhat weaker than for GDP. For example, at short-horizons only 10% of the experts are able to outperform the best performing of the crude benchmarks which is the random walk. However, for those experts performing better than the benchmarks these relative improvements are rarely significant in a statistical sense according to either the conditional or unconditional tests of predictive ability. Indeed, some of the experts exhibit a performance that is significantly worse than all of the crude benchmarks. For example, according to the tests of conditional predictive ability at the one year horizon (H=1), the expert at the 75^{th} percentile has a score that is significantly worse than either the random walk benchmark or the Naïve Market forecast. Indeed, it is only the very best performing expert that demonstrates a statistically significant improvement relative to the naïve market benchmark at the 1-year horizon. Very similar results are found from the evaluation of the 2-year-ahead inflation densities. However, for this variable, the relative performance of experts deteriorates less with an increase in the forecast horizon.

The evaluation results in Tables 5 and 6 relate to their average performance over the sample period 1999:Q1-2011:Q1. This period masks significant variation in the macroeconomic environment and it is therefore of interest to assess also the extent to which the performance of the expert densities may have varied over time. In particular, the sample includes the Great recession which followed the financial crisis which erupted in late 2007 and intensified in 2008 and 2009. As noted in the introduction, such periods represent occasions on which one might arguably expect experts to deliver additional insights compared with the mechanical benchmarks which are purely backward looking. In particular, experts could potentially draw on additional information on the nature of the crisis and its implications for macroeconomic risks and uncertainties and incorporate these insights into their survey replies. To gain some sense of the

Figure 3: Probability Scores for SPF Density Forecasts: Rolling windows
(*Ranked Probability Scores relative to the best performing benchmarks, 4 quarter moving average*)



possible time variation in the performance of the expert densities relative to the benchmarks, Figure 3 reports the RPS based on a rolling window of four quarters relative to the best performing of the benchmarks (i.e. the uniform in the case of GDP and the random walk in the case of inflation). To capture the heterogeneity in performance across individuals this moving window is presented for the median score as well as the 10th and 90th percentiles from the distribution of individual replies. Both charts highlight some significant time variation in relative performance for both variables and both horizons. In the case of GDP, there is some evidence that the performance of the densities relative to the uniform benchmark did improve during the recent recession. At short horizons (H=1), the relative scores deteriorated somewhat in late 2008 but improved substantially during the course of 2009. Moreover, during this latter period, the improvement is such that it resulted in as much as 90% of the expert scores below unity. However, there is no big "financial crisis effect" in the sense that the improvement in density

scores is similar to a previous improvement in 2004 and 2005. Moreover, at the longer horizon (H=2) there is no corresponding evidence of a significant relative improvement in the expert densities during this time. In the case of inflation, there is also some evidence of an improvement in density performance relative to the random walk benchmark during the period influenced by the crisis. Once again this is only evident for the short horizon forecasts – at the two-year horizon performance deteriorates significantly. However, even at the short horizon, for inflation the improvement is more limited with the median score remaining above unity during both 2008 and 2009. More recently, in 2010, there is a large deterioration in the performance of the expert densities particularly at the longer horizons. Overall, the evidence reported in these charts suggests some modest scope for the expert densities to improve their performance relative to crude benchmarks during the great recession of 2008 and 2009. This evidence appears stronger for GDP than for inflation and is only observable at shorter horizons.

5. DISTRIBUTIONAL FEATURES AND DENSITY PERFORMANCE

In the previous section, we have explored the relative performance of the subjective probability forecasts of the macroeconomists participating in the ECB SPF. In this section, we attempt to shed light on the main "sources" of this performance. In the case of the RPS, the density scores depend on the entire set of predicted probabilities embodied by the survey responses and not just the probability assigned to the actual outcome. We thus examine the impact on density scores of the location of the predictive densities but also of additional features of the distributions such as their spread, skew or the amount of probability mass located in their tails. Indeed, in addressing the question in the title of this paper, it is of particular interest to study the impact of these higher order distributional characteristics on individual performance both over time and across individuals. Such a mode of analysis, responds to a clear need, as expressed in Leeper (2003), to generate empirical evidence concerning the optimality and quality of information that is contained in such density forecast features. To address this need, we propose to exploit the full panel structure of the SPF data.[28] We therefore consider the relationship between individual density scores and these key distributional properties via a reduced form linear panel model of the form given in (5.1).

$$L_{i,t+\tau} = \alpha_t + X'_{i,t+\tau}\beta + \varepsilon_{i,t+\tau} \qquad (5.1)$$

[28] The application of panel techniques can help enhance the power of statistical tests aimed at evaluating macroeconomic risk forecasts. As discussed in Knüppel and Schulterfrankenfeld (2011), such tests often suffer from low to moderate power given the small sample sizes that typically characterise macroeconomic applications.

In (5.1) $L_{i,t+\tau}$ refers to the RPS of forecaster *i* given the outcome for the target variable in period $t+\tau$ and $X'_{i,t+\tau}$ is a vector of regressors capturing density features such as location, variance, skew and kurtosis that may be embodied in the subjective probability forecasts. A key advantage of such an approach is that it can identify jointly the role of key distributional features, rather than focussing individually on particular parameters. Clearly, one would expect that a large share of a given density forecast's performance will tend to be determined by its location, i.e. the point or mean forecast of the expert relative to the outcome. Controlling for this effect, however, in investigating the relative degree of skill embodied in the overall predictive distribution, it is of interest to consider the impact on the overall density forecast score of other higher dimension distributional features such as the variance, the skew or the degree of probability mass allocated by respondents to the tails of their subjective distribution. Hence, by controlling for point forecast accuracy, our analysis also offers insight into the relative information content of individual density forecasts compared to the point forecasts that are also reported as part of the ECB SPF. Below we explore these inter-linkages from both a cross section and full panel perspective.

5.1 Cross sectional evidence

A first perspective on the relationship between distributional characteristics and density forecast performance is to consider only the evidence in the cross section, averaging over time the dependent and independent variables in (5.1). This suggests a cross sectional regression of the form given in (5.2) below.

$$L_i = \alpha + \beta_1 \left| y - \mu_i \right| + \beta_2 \sigma_i^{\,2} + \beta_3 \gamma_i + \beta_4 \zeta_i + \varepsilon_i \qquad (5.2)$$

In equation (5.2), L_i refers to the sample average score of each individual density forecast, $|\, y - \mu_i \,|$ to the corresponding individual sample mean absolute forecast error, σ_i^2 to the sample average variance of forecaster *i*'s density forecasts, γ_i to the sample average *absolute* skew, and ζ_i to the sample average level of kurtosis.[29] (5.2) thus offers some insight into understanding the relevance of distributional features in systematically explaining density forecast performance across forecasters. *A priori* one would anticipate $\beta_1 > 0.0$, i.e. forecasters with relatively large mean absolute forecast errors have relatively poor density forecast performance (higher scores). The remaining parameters will tend to highlight any systematic differences across forecasters in the other distributional features (i.e. not related to point forecast accuracy) but which nonetheless systematically link to density performance. For example, to the extent that some forecasters tend

[29] We focus on the absolute skew rather than the actual skew in order to focus attention on the hypothesis that more skewed distributions are associated with higher (or lower) density forecast scores.

to be over confident or neglect certain important risks which materialised over the sample, they will tend to have low predictive variances but relatively high density scores. Conversely, those

Table 7: Distribution features and density forecast performance: Cross Sectional Evidence

$$L_i = \alpha + \beta_1 \left| y - \mu_i \right| + \beta_2 \sigma_i^{\,2} + \beta_3 \gamma_i + \beta_4 \zeta_i + \varepsilon_i$$

	α	β_1	β_2	β_3	β_4	N	R^2
GDP (H=1)	-0.26*	2.02**	-0.49**	0.15	0.01	27	0.96
	(0.14)	(0.11)	(0.06)	(0.11)	(0.04)		
GDP (H=2)	-0.13	1.85**	-0.39**	0.48**	-0.03	27	0.98
	(0.11)	(0.06)	(0.04)	(0.17)	(0.04)		
Inflation (H=1)	0.04	1.53**	-0.17*	0.13	-0.01	28	0.90
	(0.08)	(0.11)	(0.08)	(0.12)	(0.03)		
Inflation (H=2)	0.03	1.55**	-0.15**	0.08	0.00	28	0.90
	(0.1)	(0.12)	(0.06)	(0.12)	(0.03)		

Note: Cross sectional regression using OLS where all variables are expressed as sample averages over the period 1999Q1-2011Q1. Coefficients with significance level lower than 5% are indicated with "**", lower than 10% with "*". In parenthesis is reported the standard deviation of the coefficient estimate rounded to two decimal points.

Table 8: Distributional features and density forecast performance: Panel regression in levels

$$L_{i,t+\tau} = \alpha + \beta_0 D_t + \beta_1 \left| y_{t+\tau} - \mu_{i,t+\tau} \right| + \beta_2 \sigma^2_{\,i,t+\tau} + \beta_3 \gamma_{i,t+\tau} + \beta_4 \zeta_{i,t+\tau} + \varepsilon_{i,t+\tau}$$

	α	β_1	β_2	β_3	β_4	N*T	R^2
GDP (H=1)	0.32**	1.79**	-0.46**	0.01	0.01	1269	0.98
	(0.04)	(0.02)	(0.03)	(0.02)	(0.01)		
GDP (H=2)	0.47**	1.71**	-0.55**	0.04**	0.00	1161	0.99
	(0.04)	(0.02)	(0.02)	(0.02)	(0.01)		
Inflation (H=1)	0.39**	1.61**	-0.24**	0.04**	0.01	1288	0.96
	(0.04)	(0.02)	(0.03)	(0.01)	(0.01)		
Inflation (H=2)	0.36**	1.53**	-0.12**	0.06**	0.01	1176	0.96
	(0.04)	(0.02)	(0.03)	(0.02)	(0.01)		

Note: Robust parameter estimates using feasible GLS are reported, estimated over the sample 1999Q1-2011Q1 and including a correction for serial and cross-sectional correlation. Coefficients with significance level lower than 5% are indicated with "**", lower than 10% with "*". In parenthesis is reported the standard deviation of the coefficient estimate rounded to two decimal points.

with relatively lower confidence (higher variances) may be able to improve their overall density performance, i.e. lower their density scores compared with other low variance forecasters. Under these conditions, one would anticipate a significant negative relationship between average individual variance and average individual score. Similarly, the cross sectional regression allows us address the question whether individuals with more skewed distributions or distributions with greater tail mass have any tendency to perform better that those where such features are less prevalent. This might arise if, for example, some forecasters have neglected important skew or tail risks (such as those which materialised in 2008-2009) while others do not (or at least neglect them to a lesser extent).

The results from the estimation of equation (5.2) are reported in Table 7. Overall, the cross sectional regression tends to explain a large fraction of the variation in density performance across individuals with, depending on the horizon and the variable, adjusted R^2 that range between 90% and 98%. The coefficient estimates also yields a number of clear insights. Firstly, experts with less accurate point forecasts in general also have less accurate density forecasts as indicated by a significant and positive β_1 for each variable and at each horizon. Another systematic feature contained in the cross-section regression is a sizeable and statistically significant negative coefficient on the variance of the predictive densities. We would interpret this finding, which is observed for both inflation and growth and at both horizons, as providing further evidence on the role of neglected risks as highlighted previously in the discussion of Table 7. In particular, forecasters with relatively high (low) variance in their predictive densities tend to have systematically better (worse) density forecast performance. This suggests that many forecasters in the panel are operating in a region where density score can be improved by increasing predictive variance.[30] Looking in more detail at the results in Table 7, the cross section regression suggests such overconfidence is somewhat stronger for the short horizon density forecasts and is evident for both GDP growth and inflation. Turning to the remaining parameter estimates from the cross section regression, we find that forecasters with more skewed distributions (in absolute terms) tend – if anything - to do worse on average. This is indicated by an estimated positive value for β_3, although it is generally not significantly different from zero. Such a finding would cast some doubt on the "information value" of the skew assessments embodied in expert predictions, a result which compares closely with results in Knüppel and

[30] To the extent that the scoring rules used penalise symmetrically positive and negative deviations from the true but unknown variance, and if some forecasters underestimated variance while others overestimated it (i.e. there was no tendency to underestimate it on average), one would not anticipate any systematic relationship between density score and density variance in the cross-section.

Schulterfrankenfeld (2011) who report no conclusive evidence for a systematic connection between risk assessments of central banks and their forecast errors. Indeed, for longer horizon forecasts for growth, our results document a statistically significant deterioration of performance for experts with more skewed distributions compared with that of experts with less skewed distributions. Lastly, although increased probability mass in the tails is sometimes associated with an improvement in individual performance (β_4<0.0), this effect is never statistically significant. The import of this result is that those forecasters with higher probability mass in the tails of their distributions do not achieve a notable improvement in predictive performance.

5.2 Panel evidence

Another perspective on the link between distributional characteristics and density forecast performance can be obtained by estimating a full panel version of (5.2). The panel analogue of equation (5.2) is given by (5.3) below.

$$L_{i,t+\tau} = \alpha + \beta_0 D_t + \beta_1 \left| y_{t+\tau} - \mu_{i,t+\tau} \right| + \beta_2 \sigma^2{}_{i,t+\tau} + \beta_3 \gamma_{i,t+\tau} + \beta_4 \zeta_{i,t+\tau} + \varepsilon_{i,t+\tau} \tag{5.3}$$

As equation (5.3) captures the impact of changes in distributional features over time - and not just on average in the cross section - it is important to control for other factors which may vary over time but which are fixed across forecasters. A likely such factor, for example, maybe changes in the nature and volatility of the business cycle itself which impacts on all forecasters' performance. We therefore estimate equation (5.3) using a dummy variable (D_t) for time fixed effects to capture the role of these factors. Another likely factor that may impact the efficiency of the estimates in our panel regression is the possibility of serial correlation in the errors such that $E[\varepsilon_{i,t}, \varepsilon_{i,t+j}] \neq 0 \quad for \quad j = 1,..\tau \quad and \quad i = 1,...N$. We thus also control for possible correlation in the errors of the panel regression that may result from the multi-period nature of the forecast horizon and the quarterly frequency of the survey and estimate (5.3) using a Feasible GLS (FGLS) procedure.[31] Under this procedure, and using the notation of equation (5.1), we first estimate the equation residuals by estimating (5.3) by OLS to derive estimates of the error correlation. From the residuals of the OLS regression, we then construct a new estimate of the error variance covariance matrix ($\hat{\Omega}$) for our panel regression allowing for auto correlation of up to order τ.[32] The estimated parameters of the panel regression and their associated standard errors

[31] See Wooldridge (2002).
[32] In practice autocorrelations greater than 4 tend to be very small and are therefore discarded in the FGLS procedure.

are then given by $\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}L)$ and $Var[\hat{\beta}_{FGLS}] = (X'\hat{\Omega}^{-1}X)^{-1}$.[33] The results of this estimation are reported in Table 8. In general, the findings from the panel regression are very consistent with the results of the cross-section regression described previously with $\beta_1 > 0$ and $\beta_2 < 0$ for each variable and each horizon. The finding of $\beta_2 < 0$, which represents the average impact of changes in variance within the full panel, again confirms the tendency for increases in variance to be associated with improved density forecast performance. We would interpret the finding of a negative relation between density forecast variance and density score, as directly linked to the overconfidence in the sample of forecasters as discussed previously. More specifically, controlling for common factors via the fixed effects and also for the impact of point forecast accuracy on the density scores, the panel regression suggests that increases in variance would systematically yield lower scores (or improved performance). Conversely, a decrease in variance is systematically associated with higher scores and worse performance. The panel results also indicate that the higher moment features such as skew and heightened tail probability either have no significant impact on performance or result in a significant deterioration in performance. Increases in distributional skew (in absolute terms), for example, are invariably associated with a higher density score – an effect which is statistically significant for growth (H=2) and inflation (H=1 and H= 2). Fluctuations in tail risk also have no noticeable impact on density performance for any variable or at any horizon. Overall, therefore, these panel results again point to a somewhat limited ability of macroeconomists to improve their density forecast performance by manipulating such higher moment features over time.

As an additional check on the robustness of the above panel regression results, we estimate a version of equation (5.3) in first differences and consider separately the effect of increases and decreases in higher moment distributional features. A panel regression in first differences, also controlling for factors which are common across forecasters using time dummies, may help better identify individual forecaster skill. For example, to the extent that the variance of the true density has varied over time, e.g. as a result of increasing or decreasing macroeconomic uncertainty, skilled experts would be able to improve their density forecast performance by varying their predictive variances in line with this. Such skill would imply that both positive and negative change in these distributional features could result in an improvement in density forecast

[33] We also considered possible correlation in the errors across individual forecasters that might arise due to common aggregate shocks, in line with the panel analysis of point forecasts in Keane and Runkle (1990). However, we did not find these correlations to be important. A likely explanation for this finding is that the impact of common shocks is adequately captured through the inclusion of the fixed effects time dummies.

performance, as reflected in lower density forecast scores. This suggests consideration of the panel regression model (5.4) below.

$$\Delta L_{i,t+\tau} = \alpha + \beta_0 D_t + \beta_1 \Delta \left| y_{t+\tau} - \mu_{i,t+\tau} \right| + \beta_2 (\Delta \sigma^2_{i,t+\tau})^+ + \beta_3 (\Delta \sigma^2_{i,t+\tau})^- + \beta_4 \Delta \gamma^+_{i,t+\tau} + \beta_5 \Delta \gamma^-_{i,t+\tau} + \beta_6 \Delta \zeta^+_{i,t+\tau} + \beta_7 \Delta \zeta^-_{i,t+\tau} + \varepsilon_{i,t+\tau} \tag{5.4}$$

In (5.4) $\Delta L_{i,t+\tau}$ represents the change in the density score between period $t+\tau$ and $t+\tau-1$. For each of the regressors corresponding to this, we construct a dichotomous variable which takes on the value of the distributional characteristic depending on whether the change in that characteristic is positive or negative and takes a value of zero otherwise, i.e.

$$(\Delta \sigma^2_{i,t+\tau})^+ = \begin{Bmatrix} \Delta \sigma^2_{i,t+\tau} & if\ \Delta \sigma^2_{i,t+\tau} > 0 \\ 0 & otherwise \end{Bmatrix}$$

$$(\Delta \sigma^2_{i,t+\tau})^- = \begin{Bmatrix} \Delta \sigma^2_{i,t+\tau} & if\ \Delta \sigma^2_{i,t+\tau} < 0 \\ 0 & otherwise \end{Bmatrix}$$

$$\Delta \gamma_{i,t+\tau}{}^+ = \begin{Bmatrix} \Delta \gamma_{i,t+\tau} & if\ \Delta \gamma_{i,t+\tau} > 0 \\ 0 & otherwise \end{Bmatrix}$$

$$\Delta \gamma_{i,t+\tau}{}^- = \begin{Bmatrix} \Delta \gamma_{i,t+\tau} & if\ \Delta \gamma_{i,t+\tau} < 0 \\ 0 & otherwise \end{Bmatrix}$$

$$\Delta \zeta_{i,t+\tau}{}^+ = \begin{Bmatrix} \Delta \zeta_{i,t+\tau} & if\ \Delta \zeta_{i,t+\tau} > 0 \\ 0 & otherwise \end{Bmatrix}$$

$$\Delta \zeta_{i,t+\tau}{}^- = \begin{Bmatrix} \Delta \zeta_{i,t+\tau} & if\ \Delta \zeta_{i,t+\tau} < 0 \\ 0 & otherwise \end{Bmatrix}$$

To the extent that the subjective forecasts of macroeconomic experts embody "higher dimensional skill" as discussed above, one would anticipate $\beta_i < 0$ for $i = 2$, 4 and 6 and $\beta_i > 0$ for $i =$ 3, 5 and 7. Again, in examining the evidence in favour of such forecaster skill, we control for the overall location of the predictive densities using the mean absolute error and time dummies to capture those factors impacting on predictive performance that can vary over time but are common across experts.

Table 9 reports the estimation results for equation (5.4). The results again tend to confirm the previous conclusions. In particular positive changes in the variance of the predictive distributions tend to be associated with lower density scores or improved performance. Given that the estimation is in first differences, the finding of $\beta_2 < 0$ suggests that experts can indeed identify meaningful increases overtime in the "true" but unobserved uncertainty. No similar findings are, however, evident for decreases in predictive variances which tend to be systematically associated with worse performance ($\beta_3 < 0$) since the product of coefficient and the negative regressor result in a higher score. Turning to the other higher moment features, the estimates for β_4 and β_6 in these panel regressions are generally positive, implying that whenever expert densities become more skewed or fatter tailed, this is invariably associated with a worse density forecast performance or, at best, no significant impact on performance. In contrast, the panel does yield some evidence that reductions in the overall level of skew can help improve performance ($\beta_5 > 0.0$), a result which is significant for the inflation densities at short-horizons (H=1). Similarly, reductions in the degree of probability mass assigned to the tails are associated with a small though generally insignificant deterioration in performance ($\beta_7 > 0.0$).

Table 9: Distributional features and density forecast performance: Panel regression in 1st differences with asymmetries

$$\Delta L_{i,t+\tau} = \alpha + \beta_0 D_t + \beta_1 \Delta \left| y_{t+\tau} - \mu_{i,t+\tau} \right| + \beta_2 (\Delta \sigma^2_{i,t+\tau})^+ + \beta_3 (\Delta \sigma^2_{i,t+\tau})^- + \beta_4 \Delta \gamma^+_{i,t+\tau} + \beta_5 \Delta \gamma^-_{i,t+\tau} + \beta_6 \Delta \zeta^+_{i,t+\tau} + \beta_7 \Delta \zeta^-_{i,t+\tau} + \varepsilon_{i,t+\tau}$$

	α	β_1	β_2	β_3	β_4	β_5	β_6	β_7	N*T	R^2
					Δ Ranked Probability Score					
GDP (H=1)	0.63**	1.78**	-0.34**	-0.5**	0.00	-0.01	0.00	0.01	1260	0.98
	(0.04)	(0.02)	(0.03)	(0.04)	(0.02)	(0.02)	(0.01)	(0.01)		
GDP (H=2)	0.8**	1.69**	-0.51**	-0.66**	0.03	0.04	0.00	0.00	1134	0.99
	(0.04)	(0.02)	(0.04)	(0.05)	(0.02)	(0.02)	(0.01)	(0.01)		
Inflation (H=1)	0.21**	1.59**	-0.22**	-0.15**	0.03**	0.05**	0.01*	0.01	1260	0.96
	(0.04)	(0.02)	(0.04)	(0.05)	(0.01)	(0.01)	(0.01)	(0.01)		
Inflation (H=2)	0.15**	1.56**	0.08**	-0.18**	0.04**	0.03	0.01*	0.01	1148	0.96
	(0.04)	(0.02)	(0.04)	(0.05)	(0.02)	(0.02)	(0.01)	(0.01)		

Note: Parameter estimates using feasible GLS are reported, estimated over the period 1999Q1-2011Q1 and including a correction for serial correlation. Coefficients with significance level lower than 5% are indicated with "**", lower than 10% with "*". In parenthesis is reported the standard deviation of the coefficient estimate rounded to two decimal points.

6. CONCLUSIONS

The availability of reliable and informative probabilistic information on the future state of the economy is increasingly recognised as a crucial ingredient in successful macroeconomic policy making. This holds not just for the traditional domains of monetary and fiscal policy but also for new priority areas such as financial stability analysis and macro prudential supervision. In this paper we study the quantitative information actually provided by professional macroeconomists in their replies to a quarterly survey collecting their density forecasts for euro area macroeconomic outcomes and conducted since the launch of the euro. The study of such "real forecasts", as collected in surveys, contrasts with much applied forecasting literature which focuses on the "pseudo" real-time predictions of specific models which have often not been available for actual decision making. A key strength of this analysis is the exploitation of micro data and the use of evaluation criteria which examine the entire predictive densities, including an evaluation of the relative impact of density features such as location, spread, skew and tail risk on density forecast performance. Such a mode of analysis, responds to a clear need, as expressed in Leeper (2003), to generate empirical evidence concerning the optimality and quality of information that is contained in such density forecasts.

Our study proceeded along several lines of enquiry, highlighting a number of important finding. To begin with, we compare the predictive performance of the surveyed expert densities to a set of simple benchmarks which are intended to mimic i) an agnostic reply to the survey, ii) a reply based on a basic knowledge of statistical theory or iii) information that was publicly available at the time the survey was carried out. Our results uncover considerable heterogeneity in individual density performance with the performance of experts differing also depending on the forecast variable or the forecast horizon under consideration. The predictive ability of some experts is strongest for GDP growth where, at shorter horizons of one year, 75% of all experts perform better than the best performing benchmark. However, rarely, are these quantitative improvements significant in a statistical sense. Moreover, at longer forecasting horizons of two years, the superior risk assessment of the best performing experts for GDP growth is lost. Indeed, for GDP growth at this longer horizon, an agnostic uniform benchmark performs better than any of the experts in the panel! In the case of inflation forecasts, we uncover again considerable heterogeneity in the performance of expert densities. However, for this variable, although the relative performance of experts' inflation densities deteriorates less with an increase in the forecast horizon, only 10% of experts are able to outperform the best performing benchmark. Once again, however, for those forecasters that do so, these relative improvements are rarely

significant in a statistical sense. Indeed, several of the experts exhibit a performance that is statistically worse than the crude benchmarks. Overall, our findings imply some limitations on the information content of expert density forecasts. Given the inclusion in our sample of the period of macroeconomic volatility associated with the recent financial crisis, where expert judgement might have been anticipated to yield significant marginal information value compared with very mechanical assessments, we have also examined whether the performance of expert densities changes during such key phases of the business cycle. Indeed, our overall evidence suggests some scope for many experts to improve their performance during such periods, including during the Great Recession of 2009. This is stronger for GDP growth densities than it is for inflation, although it is only observable at relatively short horizons of up to a year.

Furthermore, given that our evaluation is based on the entire predictive density, we are able to take it one step further and shed light on how the features of individual expert densities impact on their overall performance. To do so, we propose a panel regression linking density forecast performance to key parameters of the forecast densities capturing their location, spread, skew and tail risk. We can thus exploit the micro evidence in the SPF dataset and thereby alleviate the problem of relatively low power which often plagues density evaluation in a macroeconomic context. Our analysis suggests clear evidence of overconfidence or neglected risks in the probability assessments of professional macroeconomists, as reflected also in frequent occurrences of events which are assigned a zero probability. In line with this, we report evidence to suggest that many experts are operating in a way where density performance could be systematically improved by correcting a downward bias in their predictive variances. Aside from this shortcoming in second moment characteristics of the individual densities, other higher moment features, such as skew or changes in the degree of probability mass given to the tails of the predictive distributions tend - as a rule - not to contribute significantly to improvements in individual density forecast performance.

The above findings should be of interest to both producers (i.e. survey participants) as well as end users of such density forecasts. In particular, experts need to reassess the degree of confidence they have embodied in their probability assessments. Similarly, forecast users need to bear in mind the limited information value of some expert predictions and the large heterogeneity that exists in predictive performance of macroeconomic experts at the individual level. Indeed, while some experts perform poorly, others have tended to outperform *both* mechanical benchmarks *and* the equal weighted linear opinion pool which aggregates all expert opinions. An important area

for future research with such survey data will be to explore in more detail this heterogeneity and, in particular, whether the information content of such surveys for forecast users can be enhanced through more optimal aggregation of the constituent expert opinions.

References

Ang, A., G. Bekaert, and M. Wie (2007), Do macro variables, asset markets or surveys forecast inflation better?, *Journal of Monetary Economics*, 54(4), 1163-1212

Bowles, C., R. Friz, V. Genre, G. Kenny, A. Meyler and T. Rautanen (2010), An evaluation of the growth and unemployment rate forecasts in the ECB SPF, *Journal of Business Cycle Measurement and Analysis*, Vol. 2010, Issue 2, 63-90

Bernardo, J. M. (1979), "Expected Information as Expected Utility," *The Annals of Statistics*, 7(3), 686–690

Boero, G., J. Smith and K. F. Wallis (2011), Scoring rules and survey density forecasts, *International Journal of Forecasting*, 27(2), April-June, 379-393

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1-3

Capistrán, C. and A.Timmermann (2009), Forecast Combination with Entry and Exit of Experts, *Journal of Business and Economic Statistics*, 27(4), 428-440

Carroll, C.D. (2003), Macroeconomic Expectations of Households and Professional Forecasters, *Quarterly Journal of Economics* 118(1), 269-298.

Casillas-Olvera, G. and D.A.Bessler(2006), Probability forecasting and central bank accountability. *Journal of Policy Modelling*, 28(2), 223-234

Clark, T. E. (2009), Real-time Density Forecasts from Bayesian Vector AutoRegressions with Stochastic Volatility, *Journal of Business and Economic Statistics*, 29(3), 327-341

Clements, M. (2010), Explanations of inconsistencies in survey respondent's forecasts, *European Economic Review*, 54(4), 536-549

Dawid, A. P. (1984), Statistical Theory: The Prequential Approach, *Journal of the Royal Statistical Society*, 147, 278-290

Diebold, F. X., and R. S. Mariano (1995), Comparing Predictive Accuracy, *Journal of Business and Economic Statistics*, 13(3), 253-63

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

Diebold, F. X., Tay, A. S. and K. F. Wallis (1999), Evaluating density forecasts of inflation: the Survey of Professional Forecasters, in Engle R. and White H. (eds.), Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W. J. Granger, *Oxford University Press*, Oxford

Dowd, K. (2008), The GDP Fancharts: An Empirical Evaluation, *National Institute Economic Review*, No. 203, January 2008

Engelberg, J., C.F. Manski, and J. Williams (2009), Comparing the point predictions and subjective probability distributions of professional forecasters, *Journal of Business and Economic Statistics* 27(1), 30-41

Epstein, E.S. (1969), A scoring system for probability forecasts of ranked categories, *Journal of Applied Meteorology*, 8, 985-987

Gennaioli, N. and A. Shleifer (2010), What comes to mind, *Quarterly Journal of Economics*, 125(4), November, 1399-1433

Gennaioli, N., A. Shleifer and R. Vishny (2011), Neglected Risks, Financial Innovation, and Financial Fragility, forthcoming, *Journal of Financial Economics*

Genre, V., G. Kenny, A. Meyler and A. Timmermann (2010), Combining the Forecasts in the ECB SPF: Can anything beat the simple average?, *ECB Working Paper* No. 1277

Geweke, J. and G. Amisano (2011), Optimal prediction pools, forthcoming, *Journal of Econometrics*, also published as Working Paper No.1017, European Central Bank.

Ghysels, E. and J. H. Wright (2009), Forecasting professional forecasters, *Journal of Business and Economic Statistics*, 27(4), 504-516

Giacomini, R. and H. White (2006), Test of conditional predictive ability, *Econometrica*, 74(6), 1545-1578

Giordani, P., and P. Soderlind (2003), Inflation Forecast Uncertainty, *European Economic Review*, 47(6), 1037–1059.

Giordani, P., and P. Söderlind (2006), Is there evidence of pessimism and doubt in subjective distributions? Implications for the equity premium puzzle, *Journal of Economic Dynamics and Control*, 30, 1027-1043

Gneiting, T. and A. E Raftery (2007), Strictly Proper Scoring Rules: Prediction and Estimation, *Journal of the American Statistical Association*, March 2007, 102(477) ,359-378

Good, I.J. (1952), Rational decisions, *Journal of the Royal Statistical Society*, 14(1), 107-114

Hall, S.G. and J. Mitchell, (2007), Combining density forecasts, *International Journal of Forecasting*, 23(1), 1-13

Kahneman, D and A. Tversky (2000), Choices, Values and Frames, *Cambridge University Press*, Russel Sage Foundation

Keane, M. P. and D. E. Runkle (1990), Testing the rationality of price forecasts: new evidence from panel data, *American Economic Review*, 80(4), September 1990, 714-735

Kilian, L., and S. Manganelli, (2008), The Central Banker as a Risk Manager: Estimating the Federal Reserve's Preferences under Greenspan, *Journal of Money, Credit and Banking* 40(6), 1103-1129

Knüppel, M. and G. Schulterfrankenfeld (2011), How informative are central bank assessments of macroeconomic risks, *Deutsche Bundesbank Discussion Paper Series* 1, Economic Studies, No. 13/2011

Leeper, E. M. (2003), An Inflation reports Report, *NBER Working Paper* No. 10089, November 2003.

Mitchell, J. and K. Wallis (2011) Evaluating density forecasts: Forecast combinations and model mixtures, calibration and sharpness, *Journal of Applied Econometrics*, 26(6), 1023-1040

Rich, R. and J. Tracy (2010), The relationship among expected inflation, disagreement, and uncertainty: Evidence from matched point and density forecasts, *The Review of Economics and Statistics*, 92(1), 200-207.

Rosenblatt, M. (1952), Remarks on multivariate Transformations, *The Annals of Mathematical Statistics*, 23(3), 470-472

Roulston, M. S., and L. A.Smith, (2002), Evaluating Probabilistic Forecasts Using Information Theory, *Monthly Weather Review*, 130(6), 1653–1660

Tay, A.S. and K.F. Wallis, (2000), Density forecasting: a survey. *Journal of Forecasting*, 19, 235-254. Reprinted in *A Companion to Economic Forecasting* (M.P. Clements and D.F. Hendry, eds.), 45-68. Oxford: Blackwell, 2002

Wallis, K. F. (2003), Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts, *International Journal of Forecasting*, 19(2), 165–175

Wallis, K. F. (2005), Combining Density and Interval Forecasts: A Modest Proposal, *Oxford Bulletin of Economics and Statistics*, 67(2005), 983-994

Wooldridge, J. (2002), Econometric Analysis of Cross Section and Panel Data, 1st edition, *The MIT Press*

Zarnowitz, V. (1969), The new ASA-NBER survey of forecasts by economic statisticians, *American Statistician*, 23(1), 12-16.

Zarnowitz, V. and L. A. Lambros (1987), Consensus and Uncertainty in Economic Prediction, *Journal of Political Economy*, 95(3), 591-621

Lahiri, K., C. Teigland, and M..Zaporowski, (1988). 'Interest rates and the subjective probability distribution of inflation forecasts', *Journal of Money, Credit, and Banking*, Vol. 20, pp. 233–248.