EUROPEAN CENTRAL BANK

EUROSYSTEM

WORKING PAPER SERIES

NO 1540 / APRIL 2013

CAN MACROECONOMISTS FORECAST RISK?

EVENT-BASED EVIDENCE FROM THE EURO AREA SPF

Geoff Kenny, Thomas Kostka
and Federico Masera

In 2013 all ECB publications feature a motif taken from the €5 banknote.

NOTE: This Working Paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Acknowledgements
The authors would like to thank Domenico Giannone, Kajal Lahiri, Tom Stark, Jonathan Wright and Ken Wallis for useful comments on this work as well as other participants at seminars in the Federal Reserve Bank of Philadelphia, 29 June 2012, the 32nd International Symposium on Forecasting, Boston 24-28 June 2012, the ECB Workshop on SPF related research on 26 November 2012 and the joint JCERESRI International Conference, Tokyo, 21 February 2013. We are also particularly grateful for many helpful comments and suggestions from an anonymous referee from the ECB working paper series. Any errors are the sole responsibility of the authors. The opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the ECB or the Eurosystem.

Geoff Kenny
European Central Bank; e-mail: geoff.kenny@ecb.europa.eu

Thomas Kostka
European Central Bank; e-mail: thomas.kostka@ecb.europa.eu

Federico Masera
Universidad Carlos III de Madrid

Address	Kaiserstrasse 29, 60311 Frankfurt am Main, Germany
Postal address	Postfach 16 03 19, 60066 Frankfurt am Main, Germany
Telephone	+49 69 1344 0
Internet	http://www.ecb.europa.eu
Fax	+49 69 1344 6000

ISSN	1725-2806 (online)
EU Catalogue No	QB-AR-13-037-EN-N (online)

Abstract

We propose methods to evaluate the risk assessments collected as part of the ECB Survey of Professional Forecasters (SPF). Our approach focuses on direction-of-change predictions as well as the prediction of relatively more extreme macroeconomic outcomes located in the upper and lower regions of the predictive densities. For inflation and GDP growth, we find such surveyed densities are informative about future direction of change. Regarding more extreme high and low outcome events, the surveys are really only informative about GDP growth outcomes and at short-horizons. The upper and lower regions of the predictive densities for inflation are much less informative.

Keywords: probability forecasts, forecast evaluation, calibration error, Survey of Professional Forecasters

JEL: C22, C53

NON-TECHNICAL SUMMARY

In this paper we attempt to shed further light on the information content of the density forecasts of macroeconomists collected in surveys. Such surveys represent a well-documented component of the toolkit available to Central Banks, including the US Federal Reserve and the ECB, as well as other policy makers, when reviewing the economic outlook and its associated risks. While most approaches to density forecast evaluation provide information on the predictive performance of the full density, we attempt to examine whether certain partitions of the SPF density forecasts provide any insights about the risk of key future macroeconomic events. In particular, we focus on the panel of SPF probability forecasts for three broad macroeconomic events. These include the probability of i) a relatively low outcome of below 1% for the target variables (growth and inflation), ii) a relatively high outturn of above 2% and, finally, iii) an increase in the forecast target variable compared with the level observed at the time the survey was carried out. By focusing on individual level expert risk assessments of these "events" our analysis can reveal aspects of the densities which are informative even if the overall aggregate density forecast exhibits a weak performance. Such information may, for example, link to heterogeneity in forecast producers' individual loss functions which may lead them to be particularly adept (or in-adept) at providing information on the likelihood of particular events.

The empirical approach we adopt focuses on the decomposition of the Quadratic Probability Score (QPS) which is a Mean Squared Error (MSE) type scoring function applied to probability forecasts. We assess the aggregate and individual level scores on

the basis of a decomposition of the QPS which highlights two key features of the forecasts that are relevant when assessing their information content. The first component is a measure of their *calibration* which refers to the correspondence between the predicted probabilities and the average frequency of occurrence of the event in question. The second component refers to their *resolution* which measures their ability to discriminate between times when the risk materialises and times when it does not. We illustrate tests for calibration and resolution in the SPF forecasts using a panel approach which exploits the full micro data of individual densities. We conduct the tests on the pooled panel of individual probability assessments and apply an estimation procedure that adjusts the variance of our estimators both for the potential presence of serial correlation (caused by overlapping forecast horizons), and potential cross-sectional dependence (caused by common aggregate shocks) in our dataset.

Our analysis yields a number of findings of relevance to central banks, and others, making use of SPF results to inform their decisions. In general, we have observed relatively low information content in the SPF density forecasts for the relatively high and low outcomes. This result is evident for inflation at one and two-year horizons and for GDP growth at two year horizons. Indeed we find that the mis-calibration of predictions for more extreme events is widely shared across individual forecasters. An exception is the information in the GDP densities for more extreme outcomes at relatively short-horizons where we observe greater reliability. In contrast to the overall poor performance in predicting tail events, the SPF densities appear considerably more informative concerning more central tendencies in the forecast target variable as reflected in their probabilistic assessments of its likely future directional change. This result, which is observed for both inflation and GDP densities, confirms the ability of survey participants to capture in their density forecasts normal cyclical fluctuations (e.g. mean reversion) in these macroeconomic variables. The evidence we uncover would thus support the case to monitor direction of change indicators from surveys such as the ECB's SPF. Lastly, our analysis also points to sizeable differences in density performance at an individual level and performance of the aggregate linear opinion pool which is the headline indicator used to publicly summarise the survey. In particular, expert densities appear far less informative and often more biased at the individual level. However, in line with the predictions from the forecast combination literature when individual level information is pooled into an aggregate density some notable improvement in forecast quality is observed.

1. INTRODUCTION

Recent experience with macroeconomic forecasting in an environment characterised by high levels of macroeconomic volatility has both highlighted the strong limitations to point forecasts as a sufficient basis for forward-looking policy deliberations *and* strengthened the demand for quality information on the risks surrounding the economic outlook. Indeed information from the entire predictive densities of future macroeconomic outcomes has an important theoretical justification in the decision sciences (see, for example, Tay and Wallis (2000)). Fortunately, such information is increasingly available in practice from different sources and often features in public discussions of the economic outlook. One such source is the Survey of Professional Forecasters (SPF) conducted by the European Central Bank (ECB) on a quarterly basis since the launch of the single currency in January 1999. Similarly, the Federal Reserve Bank of Philadelphia has an even longer tradition of collecting information on macroeconomists' assessments of future macroeconomic risks via its SPF, while the Bank of England's well known fan chart provides information on future macroeconomic risks reflecting the views of its monetary policy committee. Indeed, the Bank of England also undertakes a Survey of External Forecasters, where density forecasts are collected and aggregated in order to provide a rich probabilistic interpretation of the economic outlook.

A large literature has developed, in particular, around the density forecasts from the US survey of professional forecasters reflecting its long established track record. Diebold, Tay and Wallis (1998) employ the probability integral transform to assess the inflation densities in the US SPF. More recently, Giordani and Söderlind (2006) explore the possible role of surveyed densities in explaining the equity premium puzzle using US

SPF data, while Engelberg, Manski and Williams (2009) and Clements (2010) compare the point predictions of professional forecasters with their subjective probability distributions. Another important study is Clements (2006) who proposes techniques to evaluate forecast probabilities of *events* extracted from surveyed expert densities from the US SPF. Many of these studies point to possible shortcomings in private sector density forecasts, including evidence of excessive confidence and inattentiveness in updating probabilities in response to new information. In line with this, Lahiri and Wang (2007) find that the density forecasts of professional macroeconomists in the US SPF are informative but generally only at very short horizons. In two comparative studies, Boero, Smith and Wallis (2011) and Casillas Olvera and Bessler (2006) compare the Bank of England density forecasts with those of private experts. In general, both studies suggest that the expert densities outperform the official central bank forecasts although not significantly so. In the case of the SPF for the euro area, Kenny, Kostka and Masera (2011) have compared the overall accuracy of the individual level density forecasts from the ECB SPF with a set of simple benchmark forecasts. They find considerable heterogeneity in the performance of the surveyed densities at the individual level, with a large fraction of experts unable to outperform crude benchmark alternatives especially at longer horizons. The study by Knüppel and Schulterfrankenfeld (2012) has focussed on the evaluation of the density forecasts produced and published by central banks. These authors test empirically the optimality of measures of skew extracted from central bank density forecasts, finding that they have little systematic information content.

In this study we adopt a perspective similar to Clements (2006) and attempt to examine whether the SPF density forecasts provide any insights about the risk of key future macroeconomic events. That study focused on the conditional efficiency of aggregate probability event forecasts from the US SPF and, in particular, whether they encompass a naïve "no change" prediction for the target variable. In this study, we use similar techniques but exploit the individual level density forecasts rather than the aggregation of those individual forecasts. Most approaches to density evaluation provide information on the predictive performance of the full density. While such analysis certainly provides insight on whether or not a given density forecast is informative, it says little about how it informs or whether it is more informative about particular economic outcomes than others. One approach to gaining such insights is to partition the density at a fixed point or threshold and consider only the binary set of mutually exclusive outcomes, i.e. either the outcome ($y_{t+\tau}$) is above the specified threshold or below it. For example, one might be interested in gaining information on the ability of a particular density forecast to signal risks of relatively low outcomes to a decision maker. Denoting γ as the threshold for such lower tail outcomes, one can extract from the density the forecasted probability $f_{t+\tau} =$ Prob ($y_{t+\tau} < \gamma$). Similarly using $x_{t+\tau}$ to denote the binary indicator function taking a value of unity if $y_{t+\tau} < \gamma$ and zero otherwise, one can construct probability scores or a measure of the loss for a decision maker relying on such probabilistic assessments. In our analysis, we focus on the SPF probability forecasts for three broad macroeconomic events based on such a user defined partitioning of the forecast density functions. These include the probability of i) a relatively low outcome of below 1% for the target variables (growth and inflation), ii) a relatively high outturn of above 2% and, finally, iii) an increase in the forecast target variable compared with its current level as known at the time the survey was carried out. By focussing on expert risk assessments of these "events" our analysis can reveal aspects of the densities which are informative *even if* the overall aggregate density forecast exhibits a weak performance. Such information may, for example, link to heterogeneity in forecast producers' individual loss functions which may lead them to be particularly adept (or in-adept) at providing information on the likelihood of particular events. In helping identify aspects of the density forecasts which may be most insightful or reliable, our analysis is of primary interest for users of density forecasts including both monetary policy makers and those charged with maintaining financial stability.

Our evaluation of the above event forecasts is based on a decomposition of their associated quadratic probability score due to Murphy (1973). This decomposition focusses on two key features of the forecasts that are relevant when assessing their information content. The first component is a measure of their *calibration* which refers to the correspondence between the predicted probabilities and the average frequency of occurrence of the event in question. The second component refers to their *resolution* which measures their ability to discriminate between times when the risk materialises and times when it does not. In an economic context, the Murphy decomposition has been used to evaluate probabilistic forecasts by Diebold and Rudebusch (1989), Galbraith and van Norden (2012) and Lahiri and Wang (2007) but has been much more widely and frequently applied in the statistical and meteorological forecasting literature (Murphy (1988) and Murphy and Winkler (1992)). Mitchell and Wallis (2011) also discuss tests of density forecast calibration. Our empirical analysis exploits the full micro data of individual densities collected as part of the ECB SPF. To do so, we conduct our tests on the pooled panel of individual probability assessments and apply an estimation procedure that adjusts the variance of our estimators both for the presence of overlapping forecast horizons in our dataset as well as for the role of aggregate shocks impacting on all forecasting agents jointly.

The layout of the remainder of the paper is as follows. In Section 2, we describe in more detail the evaluation framework we adopt and its application to individual level probability forecasts using panel techniques. In Section 3, we provide some background descriptive statistics on the events and SPF probability forecasts that we examine. Section 4 presents our main empirical findings, while Section 5 concludes and summarises.

2. EVALUATING EVENT FORECASTS

In contrast to point forecasts, a probability forecast for a particular event can never be said to have been either right or wrong because we never observe the "true" probability. However, when such forecasts are issued over a period of time, it is nonetheless possible to apply checks of their "external validity", i.e. evaluating their correspondence with the related outcome over time. As reviewed in Dawid (1982), a long tradition exists on testing the external validity of probability forecasts in the statistical and meteorological forecasting literature (e.g Murphy (1973), Yates (1982), Murphy and Winkler (1992) and Gneiting, Balabaoui and Raftery (2007). Such methods involve gauging the usefulness of such forecasts with respect to the observed outcome and have also been applied to economic forecasting by, among others, Berkowitz (2001), Clements (2006), Lahiri and

Wang (2007), Galbraith and Van Norden (2012), Boero, Smith and Wallis (2011), Mitchell and Wallis (2011) and are closely related to the field of interval forecasting discussed in Christoffersen (1998).[1] The approach we adopt here is very much in the spirit of Berkowitz (2001) insofar as we emphasise the evaluation of the entire distribution. However, in contrast to Berkowitz, our approach focuses on the decomposition of the Quadratic Probability Score (QPS) which is a Mean Squared Error (MSE) type scoring function applied to probability forecasts and originally suggested by Brier (1950). The latter is directly analogous to the MSE of a point forecast with the exception that the outcome variable ($x_{t+\tau}$) is a binary random variable taking a value of unity when the event occurs and zero if it does not. The QPS($f_{t+\tau}$, $x_{t+\tau}$) = E[$f_{t+\tau}$ - $x_{t+\tau}$]^2 thus provides a scoring rule which penalises forecasts ($f_{t+\tau}$) which assign a low (high) probability to events that occur (do not occur). To shed light on the attributes and validity of probability forecasts, Murphy (1972) suggested a factorisation of the QPS based on the conditional distribution of $x_{t+\tau}$ given $f_{t+\tau}$,, i.e.

$$QPS(f_{t+\tau}, x_{t+\tau}) = \sigma_x^2 + E_f[\mu_{x/f} - f]^2 - E_f[\mu_{x/f} - \mu_x]^2 \tag{2.1}$$

The first term on the right hand side of (2.1) measures the unconditional variance of the binary outcome variable which can be seen as a proxy for the difficulty of the specific forecasting situation. The second term measures the overall reliability or *calibration* error of the forecasts as the difference between the forecast probability (*f*) and expected frequency of occurrence given the forecasts ($\mu_{x|f}$). Well calibrated probability forecasts

[1] Granger and Peseran (2000) argue in favour of a closer link between decisions of forecast users and the forecast evaluation problem, stressing also the importance of predictive distributions. In this respect, the recent work of Andrade, Gyhsels and Idier (2011) highlights the value in SPF distributions by helping to identify their potential role in the central bank reaction function.

are approximately valid or “unbiased in the large” (Murphy and Epstein (1967)). All other things equal, mis-calibrated forecasts will tend to have a larger QPS. However, even perfectly calibrated forecasts can be clearly unsatisfactory if the forecaster is unable to gauge the timing of the event. The last term on the right hand side of the equation provides a measure of the *resolution* of the forecasts. Resolution contributes negatively to the QPS all other things equal. It captures the ability of forecasters to use their probability forecasts to sort individual outcomes into groups which differ from the long run or unconditional relative frequency of occurrence (μ_x). Probability forecasts with high resolution will therefore take values that are further away from the mean frequency of occurrence and closer to the zero or one extremes. Even though well calibrated forecasts are desirable, it is resolution which can give a particular probability forecasts its signalling quality and thus give it some practical usefulness. A constant probability forecast that is always equal to the relative frequency of occurrence (i.e. $f_{t+\tau} = \mu_{\mathrm{x}}$) is perfectly calibrated but it is completely uninformative from a decision makers’ perspective. High resolution is not, however, an end in itself. Given the inability to predict the future with complete certainty, there will tend to be a trade-off between the degree of forecast resolution and the calibration error. The decomposition in (2.1) is therefore not an orthogonal one and, at some point, greater resolution will tend to be associated with an overall increase in calibration error and a resulting deterioration in the QPS. The art of probability forecasting can thus be viewed as trying to minimise (2.1) by optimally managing such a trade-off between the information gain that emerges from having high resolution and the associated reduction in overall accuracy (and mis-calibration) that high resolution forecasts may ultimately introduce. Of course, the extent of this trade off will most likely differ depending on the forecasting situation, e.g. depending on the economic variable, the forecast horizon or the particular economic context.

Murphy and Winkler (1992), Galbraith and van Norden (2012), Lahiri and Wang (2007) discuss econometric regression based tests of "perfect" calibration (i.e. zero calibration error) and "zero" resolution (i.e. no skill of forecasters in sorting outcomes). Such tests are based on a generalisation of the forecast-realisation regressions originally suggested in Mincer and Zarnowitz (1969) to probability forecasts, an approach which has also been applied in the closely related literature on interval prediction (see, Christoffersen, 1998). However, to our knowledge all previous applications have ignored individual level forecasts. In a panel context with multi-period forecasts, for a given forecast horizon (τ), both tests can be constructed by regression of the outcome in period $t+\tau$ on a constant and the probability forecasts of individual i for the same period:

$$x_{t+\tau} = \alpha + \beta f_{i,t+\tau} + \varepsilon_{i,t+\tau} \tag{2.2}$$

Under the null hypothesis of perfectly calibrated forecasts, we would expect $\alpha = 0$ and β= 1. Similarly, the forecasts have zero resolution if $\beta = 0$. Ideally, for good probability forecasts, we would want to *accept* the hypothesis of perfect calibration but *reject* the restrictions implied by zero resolution. In testing these hypotheses, some attempt is needed to control for serial correlation induced by the multi-period nature of the forecast horizon, e.g. by using a correction to the standard errors of the parameters suggested by Newey and West (1987). In applying (2.2) to individual level data, however, an additional complication arises due to the role of aggregate shocks (reflecting the commonality of the outcome variable across individual forecasters) which can result in strong co-movement in forecast errors across individuals. As a result, the panel regression (2.2) will have errors that are correlated across individuals. A failure to control for the impact of aggregate shocks in such regressions will tend to bias downward the estimated standard errors and, hence, bias the tests in favour of rejecting either the null of perfect calibration or zero resolution. We therefore propose to control for the impact such serial correlation and aggregate shocks using a more generalised residual covariance matrix ($\hat{\Omega}$). As demonstrated in Keane and Runkle (1990), we can construct a covariance matrix that is consistent even in the presence of aggregate shocks, under the following two simplifying assumptions:

$$E[\varepsilon_{i,t+l}\,\varepsilon_{j,t+m}] = \left\{ \begin{array}{l} \sigma_{|l-m|} \geq 0 \quad for\ \ i = j\ \ and\ |l-m| \leq \tau \\ \delta_{|l-m|} \geq 0 \quad for\ \ i \neq j\ \ and\ |l-m| \leq \tau \\ 0 \quad otherwise \end{array} \right\} \tag{2.3}$$

Whilst $\sigma_{|l-m|}$ refers to the within individual correlations and thus allows for residual autocorrelations linked to the multi-period forecast horizon, $\delta_{|l-m|}$ additionally reflect correlations between individuals that are due to aggregate shocks. We restrict all elements of $\hat{\Omega}$ to be non-negative since only positive correlations are economically meaningful. Moreover, equation (2.3) implies that the probability forecasts' errors are not conditionally heteroskedastic and that no forecaster is systematically better than any other forecaster (i.e. both the σ_i and the δ_i are the same for each individual).[2] The corrected estimate of $\hat{\Omega}$ thus has several off-diagonal non-zero elements capturing both the time and cross sectional correlation in the residual of (2.2) and takes the form given in equation (2.4) below.

$$\hat{\Omega} = \begin{bmatrix} A & B & . & . & . & B \\ B & A & B & . & & \\ B & B & A & B & . & \\ . & . & . & & & \\ B & . & . & . & & A \end{bmatrix} \tag{2.4}$$

and where the matrices A and B are constructed as follows

$$A = \begin{bmatrix} \sigma_0 & \sigma_1 & \sigma_2 & ..\sigma_\tau & 0 & 0\,.. & 0 & 0 & 0 & 0 \\ \sigma_1 & \sigma_0 & \sigma_1 & .... & \sigma_\tau & 0\,.. & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_\tau & \sigma_{\tau-1} & \sigma_0 & \dots & \sigma_2\,.. & 0 & 0 & 0 & 0 & \\ 0 & \sigma_\tau & \sigma_0 & \dots & \sigma_1\,.. & 0 & 0 & 0 & 0 & \\ 0 & 0 & \sigma_\tau & \dots & \sigma_0\,.. & 0 & 0 & 0 & 0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & 0 & \sigma_\tau & . & . & . & ...\sigma_1 & \sigma_0 & \sigma_1 \\ 0 & 0 & 0 & 0 & 0 & \sigma_\tau & ...\sigma_3 & \sigma_2 & \sigma_1 & \sigma_0 \end{bmatrix} \tag{2.5}$$

[2] Although at first pass, this assumption is quite restrictive, it has some mixed empirical support. D'Agostino, McQuinn and Whelan (2012) offer some recent evidence for the point forecasts from the US SPF which provides partial validation for this assumption. In particular, they find limited evidence for the idea that the best forecasters are actually innately better than others, though there is evidence that a relatively small group of forecasters perform very poorly. As cited in Keane and Runkle (1990), earlier studies on US data such as McNees (1975) report similar findings. The study by Genre, Kenny, Meyler and Timmermann (2010) using the ECB SPF, on the other hand suggests that there may be differences across economic variables and horizons. For example, while they find little evidence of systematic differences in individual forecast performance for growth and the unemployment rate, in the case of short-term inflation forecasts, this latter study finds evidence of some systematic persistence in individual forecast performance which can be exploited by flexible forecast combination techniques.

$$B=\begin{bmatrix} \delta_0 & \delta_1 & \delta_2 & ..\delta_\tau & 0 & 0\,..\,0 & 0 & 0 & 0 \\ \delta_1 & \delta_0 & \delta_1 & .... & \delta_\tau & 0\,..\,0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ \delta_\tau & \delta_{\tau-1} & \delta_0 & \dots & \delta_2\,.. & 0 & 0 & 0 & 0 \\ 0 & \delta_\tau & \delta_0 & \dots & \delta_1\,.. & 0 & 0 & 0 & 0 \\ 0 & 0 & \delta_\tau & \dots & \delta_0\,.. & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & \delta_\tau & .\ .\ .\ ...\delta_1 & \delta_0 & \delta_1 \\ 0 & 0 & 0 & 0 & 0 & \delta_\tau & ...\delta_3 & \delta_2 & \delta_1 & \delta_0 \end{bmatrix}$$

(2.6)

Using (2.4) it is then possible to derive the generalised least squares estimates for α and β in the standard way and to draw inference on the values of these parameters that is robust even in the presence of aggregate shocks. In particular, we base our hypothesis tests on the Feasible Generalised Least Squares (FGLS) procedure described in Wooldridge (2002). This procedure involves first estimating (2.2) using OLS and then deriving estimates of the elements of $\hat{\Omega}$ from the associated OLS residuals under the constraints implied by (2.3). The elements of (2.4) are then computed as in (2.7) and (2.8).

$$\sigma_{|l-m|} = [N(T-|l-m|)]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T-|l-m|}\varepsilon_{i,t},\varepsilon_{i,t+|l-m|} \quad for \quad |l-m| = 0,.,.,\tau$$

(2.7)

$$\delta_{|l-m|} = [N(N-1)(T-|l-m|)]^{-1}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{t=1}^{T-|l-m|}\varepsilon_{i,t},\varepsilon_{j,t+|l-m|} \quad for \quad |l-m| = 0,.,.,\tau$$

(2.8)

Using X to denote the NT x 1 matrix containing N stacked output variables and F to denote the corresponding NT x 2 matrix for the regression constant and the stacked individual probability forecasts, the estimated regression parameters and their variance co-variance matrix are given by $\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}(X'\hat{\Omega}^{-1}F)$ and $Var[\hat{\beta}_{FGLS}] = (X'\hat{\Omega}^{-1}X)^{-1} = V$ respectively. In practice, as (2.7) and (2.8) do not ensure satisfaction of the inequality constraint in (2.3), any negative values for $\sigma_{|l-m|}$ and $\delta_{|l-m|}$ are set to be zero. The test of perfect calibration implies a joint restriction on the model's two parameters under the null hypothesis and can be implemented as a Wald test using the χ^2 distribution with 2 degrees of freedom. In the case of the test of zero resolution, we use a simple *t*-type test with N*T-2 degrees of freedom.

3. DATA: EVENT FORECASTS FROM THE SPF

In this section we provide a descriptive review of the event forecasts we evaluate using the methods described in Section 2. The complete underlying micro dataset can be downloaded at http://www.ecb.europa.eu/stats/prices/indic/forecast/html/index.en.html. Our analysis is based on the one and two-year horizon density forecasts for euro area real output growth and consumer price inflation, with these variables being measured, respectively, by euro area Gross Domestic Product (GDP) and the Harmonised Indicator of Consumer Prices (HICP) and published by Eurostat, the statistical agency of the European Union, in the first quarter of 2012.[3] The analysis is conducted using a filtered SPF dataset which excludes irregular respondents as described in Genre *et al.* (2010) and

[3] A potentially important factor impacting the evaluation of density forecasts concerns the vintage of the series used to measure the outcome variable, i.e. whether the first estimate or subsequently published revised estimates are used. Genre, Kenny, Myler and Timmermann (2013) have recently examined this issue and found little sensitivity for the evaluation of point forecasters in the case of the euro area especially for inflation which has tended to be revised only little during their evaluation sample. Future research should however broaden the analysis of this issue to the case of density forecasts.

draws on the quarterly rounds of the SPF conducted over the period 1999Q1-2011:Q3. As such the data comprises a cross sectional dimension of 24 to 26 individual forecasters depending on the particular forecast variable or horizon. This set of individual respondents represents a subset of regularly responding forecasters based on a filtering rule that excludes those forecasters who have missed more than four consecutive survey rounds. As a result, the dataset is an unbalanced panel with the precise number of time series observations varying at the individual level depending on how often a given individual has not submitted a response to the survey. More complete descriptions of the SPF dataset, including a description of its panel dimension, is given in Bowles *et al.* (2010) and Genre *et al*. (2011). Garcia (2003) provides an earlier "bird's eye" description of the ECB SPF.

Figure 1 and 2 provide a summary of the different events considered in our subsequent evaluation. In the case of GDP growth, the chart indicates four occasions during which growth exceeded the 2% threshold we use for the analysis. Conversely there were three occasions when growth fell below the lower threshold of 1%, most notably during the great recession of 2008 and 2009. Similarly, at the 1-year horizon, there were five occasions where the GDP growth outcome that emerged was higher than the current growth rate observed at the time the survey was carried out. In the case of inflation, the pattern is somewhat different with annual inflation being quite often above the 2% threshold we use for this study. Given this outcome, if the probability assessments of SPF participants are well calibrated, we might expect to observe relatively high probabilities for this event. In contrast, the below 1% outcome for inflation has occurred only once during the sample (during the 2009 downturn linked to the recent financial crisis). For aggregate analysis, there are considerable limitations when empirically testing the information content of expert probability assessments given that we observe the event only once in the sample. However, given that we employ individual level data in the subsequent analysis we are able to conduct statistical inference concerning the ability of macro economists to assess the likelihood of these infrequently occurring events. Finally, reflecting also the tendency to observe higher inflation outcomes more frequently than lower ones, the direction of change event that inflation turns out to be higher than the level persisting at the time of the survey, has also occurred quite frequently at both the one and two year forecast horizons.

To conduct our analysis, we also need to extract the cumulative probabilities related the three events that we analyse. SPF respondents submit their replies in the form of discrete histograms assigning probabilities to a set of intervals representing possible outcome ranges for the target variable. In addition, at the extremities of these histograms, the assigned probabilities relate to open intervals. To extract the event probabilities from the SPF data we make the assumption that the probabilities within a given range are uniformly distributed within that range. Without further information on the possible distributional perceptions of survey respondents, the assumption of uniformity seems the most reasonable. An alternative approach would be to fit specific continuous densities to the individual level data and derive associated event probabilities from them. However, such an approach could involve the introduction of substantial measurement error. In addition, regarding the open intervals at the edges of the histogram, these are assumed to be closed intervals of equal width to the other surveyed intervals. We have conducted some sensitivity analysis to an alternative assumption that the open intervals are twice the width of the closed intervals and found no notable impact. This is related to the properties of the survey data where in fact, at the individual level, it is often the case that either a very small or a zero probability is assigned to these open intervals.

Figure 3 and 4 present the probability forecasts for the three different events for GDP growth and inflation at both forecast horizons. The probabilities are depicted showing the median probability together with the 10th and 90th percentiles extracted from the cross section of individual surveyed densities. Also reported is the probability extracted from the aggregate SPF density, which in general is often very close to the median probability. For each chart, we also depict using shading the time periods in which the event in question actually occurred. From the charts, the direction-of-change assessments appear to correlate quite well with the actual occurrence of the events. This is the case especially for GDP at the one-year horizon but some clear correspondence between the occurrence of this event and the probability forecasts is also observed at longer horizons and also for inflation. This first visual inspection of the data suggests a less clear correspondence between the expert probability assessments for the more extreme economic outcomes represented by both the upper and lower thresholds. A good example of this is the probability assessment for low inflation. In the case of the one year horizon, the probabilities for this event (Figure 4) appear to be lagging, starting to rise only after the event in question had actually occurred. In the case of the two-year assessment, this lagging pattern is even more evident with the probabilities only starting to rise after the low inflation outcome had completely passed. The graphical evidence for other extreme outcomes also suggests relatively limited signalling power of the SPF probability assessments. An exception is perhaps the one-year ahead GDP predictions for both high and low outcomes. In the next section, we exploit the QPS decomposition and calibration and resolution tests using the full panel of probability forecasts in order to shed more robust econometric evidence on the information content of the SPF densities.

4. EVALUATION RESULTS FOR SPF EVENT FORECASTS

In this section we report the results from the evaluation of the SPF event forecasts discussed above. We first report the QPS decomposition, providing evidence of mis-calibration and signalling power (resolution). We then report more formal test of perfect calibration (unbiasedness) and zero resolution (no signalling power) using both aggregate and pooled individual level data from the SPF. Finally, we explore the heterogeneity in the SPF panel in more detail.

4.1 GDP Growth Events

Table 1 reports the QPS and its associated decomposition for each of the three GDP events at both one and two year horizons. The decomposition is based on the aggregated probabilities which average the probabilities derived from each of the individual SPF densities. The QPS statistics indicate that the SPF densities perform less well at capturing the more extreme threshold events, whilst the direction-of-change predictions perform better. For all three events at short horizons (H=1), the aggregate SPF probabilities appear close to perfectly calibrated (as indicated by a very small calibration error). They also exhibit some positive resolution, which is particularly strong for the direction of change forecasts. In contrast, the signalling information provided by the extremities of the GDP densities is smaller. At the two year horizon, there is a notable increase in the mis-calibration for the extreme event forecasts, while the calibration error for the direction of change forecast continues to be reasonably small. The latter forecast also continues to possess useful signalling information as reflected in its estimated resolution even at the longer horizon. Overall, therefore, the QPS decomposition suggests the SPF densities for GDP are most informative at short horizons and provide less reliable information about future events at the extremities. In contrast, the direction of change information appears informative, even at longer horizons. We can, however, provide more formal evidence on this using the regression based tests of perfect calibration and zero resolution described in Section 2.

Table 2 reports the results from the estimation of Equation 2.2 based on the aggregates probabilities and including a correction for serial correlation in the errors using the FGLS procedure described in Section 2. The estimation results tend to confirm the observations made above. In particular, at short horizons we are unable to reject the hypothesis of perfect calibration for all three event forecasts. Indeed, at this horizon, the parameter estimates for the low growth outcomes and the direction of change forecast are remarkably close to their predicted values under the null hypothesis of perfect calibration. Not surprisingly, therefore, the χ^2 test is unable to reject the null hypothesis of well calibrated forecasts. For the three events considered, we can also reject the null hypothesis of zero resolution. This tends to confirm the useful signalling information in the SPF densities for GDP at short horizons, even for less frequent or more extreme events. This evidence is strongest for the low growth outcomes and the direction of change forecasts, with the null of zero resolution for outcomes > 2.0% only rejected at the 5% level of significance. At longer horizons, these relatively positive findings are reversed, however. For both high and low outcomes, we can reject the null of perfect calibration. The null of zero resolution is also accepted for both high and low growth outcomes at the two year horizon. The estimates of β for these events at this longer horizon also tend to be negative, implying that the probabilities tend to fall when these more extreme events occur. Such an inverse correlation points to the relatively poor information content of the aggregate densities for such events at longer horizons. For the direction of change forecast at this longer horizon, however, we continue to accept the null of perfect calibration and reject the hypothesis of zero resolution. Hence, even at this longer horizon, econometric results tend to confirm some important information content of the SPF densities for the direction of change in GDP.

Table 3 reports the equivalent regression results for all three GDP event forecasts but based on the full unbalanced panel of individual responses and including a correction for both serial correlation and aggregate shocks (again using the FGLS estimation procedure described in Section 2). The panel results tend to confirm many of the findings observed in Table 2 using the aggregate level data. For short horizons, the hypothesis that the probability forecasts exhibit zero resolution is strongly rejected. At longer horizons, however, the probability forecast for relatively high growth outcomes exhibits no resolution while that for relatively low growth outcomes is inversely correlated with the outcome. Although our event does not corresponding to a recession in a classical sense of negative growth, this latter result for long horizons is broadly in line with Harding and Pagan (2010) who review the literature and empirical evidence on the predictability of recessions and conclude that it is very difficult to predict these events *ex ante*. One interesting feature of the panel results, which contrasts with the aggregate results, is the strong rejection of the perfect calibration hypothesis at the individual level for all three GDP events at the one-year forecast horizon. In particular, even controlling for the impact of common shocks and serial correlation in the errors of the panel regression, the estimated standard errors are such that the null of perfect calibration tends to be rejected at the individual level. This contrasting finding on the calibration of the event forecasts from the SPF at the aggregate and individual level provides some justification for the conduct of surveys such as the SPF. In particular, in line with the predictions from the density forecast combination literature surveyed recently in Timmermann (2006), when individual level information is pooled into an aggregate density forecast, some notable improvement in forecast quality may be obtained.[4] Lastly, and much more in line with the regression results based on aggregate level data, the panel estimates highlight the very poor calibration of high and low growth outcomes at longer horizons. Our empirical results therefore point to a quite dramatic deterioration in the information value of density forecasts for real output growth when the forecast horizon is extended from one to two years.

[4] The contrasting results between the aggregate and pooled estimations in Table 2 and 3 suggests significant heterogeneity in individual SPF forecasters, i.e. that the estimated parameters may differ across individual forecasters. In Section 4.3 below we provide some further evidence and discussion on the nature of this heterogeneity.

4.2 Inflation Events

Table 4 reports the QPS and its associated decomposition for each of the three inflation events at both one and two year horizons. The results compare somewhat less favourably with the previous findings for GDP. In particular, the inflation scores are as a rule higher than the corresponding GDP scores. Moreover, SPF probability forecasts for both high and low inflation outcomes show signs of mis-calibration – even at the shorter horizons. Similarly, the probability forecasts for outcomes more toward the extremes exhibit quite low signalling power as reflected in low resolution. However, once again, the direction of change forecasts appear better calibrated and exhibit positive resolution. This is mainly at the one-year horizon, however, and there is less evidence that the direction of change forecasts for inflation provide useful signals (i.e. positive resolution) at longer horizons.

Table 5 and 6 reports the econometric tests for perfect calibration and zero resolution for each of the three inflation events at both one and two-year horizons. At the aggregate level, the results for inflation mirror some of the previous findings observed for GDP. In particular, direction-of-change forecasts appear quite informative as they are better calibrated and exhibit significant positive resolution. As with GDP, this finding is again most strongly observed for the aggregate probabilities, while at the individual level the panel estimates suggest some clearer signs of mis-calibration. Predictions for more extreme outcomes exhibit signs of mis-calibration both at the aggregate and the individual level. For example, in the case of inflation, our results highlight strong evidence of mis-calibration for both high and low inflation outcomes even at short-horizons (a result which compares less favourably to the GDP growth predictions).

Moreover, as indicated by the estimated β parameters, the SPF predictions for more extreme inflation events appear to be either uncorrelated or correlate negatively with the occurrence of these events. Hence, in the case of inflation, our results would strongly suggest for users of the SPF information to exercise considerable caution when extracting information on the likelihood of more extreme events from the SPF distributions.

4.3 Heterogeneity in individual event forecasts

The preceding analysis has focused on documenting the event forecast performance of macroeconomists drawing on tests based on aggregate probability distributions or the pooled individual densities. Such an approach directly addresses the question posed in the title of this paper as it sheds light on whether or not surveyed densities are informative "in the large". However, it says little about the extent and nature of heterogeneity in performance at an individual level. To shed light on this, we have also estimated equation (2.2) for each individual in the filtered SPF panel. As an example, Figure 5 provides information on the estimated constant and slope parameter at the individual level for the case of high inflation outcome and at relatively short horizons. The Figure depicts a histogram measuring on the vertical axis the number of individual forecasters with the estimated parameters values indicated by the range of values on the horizontal axis. Also reported are the parameter estimates based on the aggregate distributions (indicated by a dotted vertical line and taken directly from Tables 2) together with the median parameter values (indicated by a solid vertical line). The histograms confirm that the relatively poor calibration of SPF forecasts for this inflation event is broadly shared across the majority of forecasters as indicated by estimates of α which are consistently above zero. Indeed all forecasters in the panel have tended to under-predict the occurrence of relatively high inflation outcomes. In terms of signalling power (resolution), the individual level parameter estimates for β suggests even greater heterogeneity with the estimates ranging from -0.5 (in the case of a single forecaster) to 1.0 in the case of the forecaster with the highest resolution. Such heterogeneity suggests that, while forecasters are generally poor at predicting such extreme inflation events, a few individuals are better able to signal inflation risks than others.[5]

Table 7 summarises the results of the individual by individual regressions for all three events, for both target variables and both forecast horizons. For the hypothesis of perfectly calibrated forecasts, the table reports the number of individuals for which the hypothesis is rejected (at the 10% level) expressed as a share of the total number of individuals in the panel. Hence, the table provides some summary information on the degree of heterogeneity in density forecast performance. Shares at the upper and lower bounds, i.e. that are close to either 0% or 100%, indicate a high level of homogeneity in forecast performance, while shares that are away from these bounds highlight some notable heterogeneity in predictive performance. The figures in Table 7 confirm that the mis-calibration of predictions for more extreme events for inflation at both horizons and for GDP growth at the longer horizon (H=2) is widely shared across individual forecasters. In particular, as indicated by rejection shares of 100%, in these instances the hypothesis of well-calibrated probability forecasts is rejected for all forecasters in our

[5] This suggests the need to investigate alternatives to the current practice of taking an equal weighted average when aggregating individual SPF replies. For example, Jore Mitchell and Vahey (2010) have identified gains from combined density forecasts that weight more highly the more informative component densities. When the number of density forecasts available is large, estimation of individual density weights can become computationally burdensome. Recently, however, Conflitti, De Mol and Giannone (2013) have proposed methods to estimate optimal combination weights and applied their method to the euro area SPF, finding some modest gains.

panel (implying a high degree of homogeneity). In some contrast, the previous findings of reasonably well calibrated direction of change forecasts for both growth and inflation, is shown to be not fully shared by all experts in the panel. For example, for the direction of change forecasts for GDP at short horizons (H= 1), the hypothesis of perfect calibration is rejected for up to 38% of individuals.

Table 7 also provides information on the resolution and signalling power of the probability forecasts at the individual level. In particular, it reports the number of individuals for which the one sided hypothesis $\beta \leq 0$ is rejected (at the 10% level) also expressed as a share of the total number of individuals. This contrasts somewhat with the two-sided t-test for zero resolution reported earlier but is an additional useful test given that it provides information on the sign of the estimated slope parameter and, hence, on the correlation between the probabilities and the event outcomes. Rejection of this hypothesis implies that the probability forecasts are informative in the sense that they have non-zero resolution *and* are positively correlated with the events occurrence. A significant degree of heterogeneity in the signalling power of GDP event forecasts is observed for relatively extreme growth outcomes above 2.0% with only 62% of all forecasters rejecting this hypothesis. For the other events and forecast horizons, these individual level results imply considerable homogeneity in terms of resolution. For example, the relatively high information content of direction of change forecasts is confirmed for nearly all forecasters for both growth and inflation and at both forecast horizons. Similarly, the relative non-informativeness of the inflation densities about more extreme inflation outcomes is widely shared across forecasters. For example, for only 4%

of forecasters (one forecaster in our panel), is the hypothesis of $\beta \leq 0$ rejected for the event that inflation would be above 2.0% (H=1 and H=2). Similarly for no forecaster in the panel, does this test indicate the densities are informative about low inflation events.

The above analysis of density forecast heterogeneity demonstrates that when such forecasts are non-informative, this poor performance tends to be a widely shared across the sample of forecasters in our panel. In contrast when we observe some information content in the density forecasts, such as is the case for direction of change forecasts for both GDP growth and inflation or for more extreme GDP outcomes but only at short horizons, such forecaster skill appears less evenly distributed across individuals. Future research might therefore consider the extent to which such heterogeneity in forecaster skill can be exploited in order to improve the usefulness of surveys such as the SPF. For example, it may be possible to enhance aggregate density performance by excluding some forecasters whose densities exhibit persistently poor calibration and/or low information content.

5. DISCUSSION AND CONCLUSIONS

In this paper we have attempted to shed further light on the information content of the density forecasts of macroeconomists collected in surveys. Such surveys represent a well-documented component of the toolkit available to Central Banks, including the US Federal Reserve and the ECB, as well as other policy makers, when reviewing the economic outlook and its associated risks. A key feature of the study has been the application of forecast evaluation methods based on a partitioning of the SPF density forecasts in order to extract probabilistic assessments for important future events that may be of interest to forecast users. The events in question include the occurrence or non-occurrence of relatively extreme outcomes (e.g. low or high growth) or qualitative assessments concerning the future direction of change in the forecast target variable (e.g. whether it will be higher or lower than its current value). Another important feature of our study is that the empirical analysis exploits the micro features of the SPF data set and controls for the impact of aggregate shocks when drawing inference concerning the information content of such probabilistic assessments.

Our analysis yields a number of findings of relevance to central banks, and others, making use of SPF results to inform their decisions. In general, we have observed relatively low information content in the SPF density forecasts for relatively high and low outcome events. This result is evident for inflation at one and two-year horizons and for GDP growth at two year horizons. Indeed we find that the mis-calibration of predictions for more extreme events is widely shared across individual forecasters. An exception is the information in the GDP densities for more “extreme” high and low outcomes at relatively short-horizons where we observe greater reliability. Overall, we would interpret these results as highlighting a need for caution when extracting information from SPF densities concerning the likelihood of more extreme events. This is unfortunate because, as highlighted in Killian and Manganelli (2008) and Andrade, Ghysels and Idier (2011), such information on tail risks is of potential use to decision makers in responding and managing macroeconomic risks. In contrast, the SPF densities appear considerably more informative concerning more central tendencies in the forecast target variable as reflected in their probabilistic assessments of its likely future directional change. This result, which is observed for both inflation and GDP densities, confirms the ability of survey participants to capture in their density forecasts normal cyclical fluctuations (e.g. mean reversion). The evidence we uncover would thus support the case to monitor direction of change indicators from surveys such as the ECB's SPF. Lastly, our analysis also points to important differences in density performance at an individual level and performance of the aggregate linear opinion pool which is the headline indicator used to publicly summarise the survey. In particular, expert densities appear far less informative and often more biased at the individual level. However, in line with the predictions from the forecast combination literature, e.g. as discussed in Geweke and Amisano (2011) or Timmermann (2006), when individual level information is pooled into an aggregate density some notable improvement in forecast quality is observed. While that implies that economists have some way to go before they could claim to be delivering reliable density predictions at the individual level, particularly for the more extreme events, it nonetheless also provides some justification for the information gain that can be achieved through the conduct and aggregation of expert replies to surveys such as the SPF.

REFERENCES

Andrade, P., E. Ghysels and J. Idier, (2011), Tails of inflation forecasts and tales of monetary policy, *mimeo*, Banque de France.

Benjamini,Y. and Y. Hochberg, (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.

Berkowitz, J. (2001), Testing density forecasts with applications to risk management, *Journal of Business and Economic Statistics*, 19(4), 465-474

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

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

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

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

Christoffersen, P. F. (1998), Evaluating interval forecasts, *International Economic Review*, 39, 841-862

Clemen, R.T (1989), Combining forecasts: A review and annotated bibliography, *International Journal of Forecasting*, 5 (4), 559-583.

Clements, M. P. (2006), Evaluating the Survey of Professional Forecasters probability distributions of expected inflation based on the derived event probability forecasts, *Empirical Economics*, 31, 49-64

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

Conflitti C, C. De Mol and D. Giannone (2012). Optimal Combination of Survey Forecasts, *ECARES Working Papers*, 2012-023, Universite Libre de Bruxelles.

D'Agostino, A., K. McQuinn and K. Whelan, (2012), Are Some Forecasters Really Better Than Others?, *Journal of Money Credit and Banking*, 44(4), 715-732

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

Diebold, F. X. and G.D. Rudebusch (1989), Scoring the leading indicators, *Journal of Business*, 64, 369–91.

Diebold, F. X., A.S. Tay 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

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

Galbraith, J. W. S. van Norden (2012), GDP and Inflation Probability Forecasts Derived from the Bank of England Fancharts, *Journal of the Royal Statistical Society,* Series A, 175(3), 713-727

Garcia, J.A. (2003), An Introduction to the ECB Survey of Professional Forecasters, *ECB Occasional Paper* No. 8, September.

Genre, V., G. Kenny, A. Meyler and A. Timmermann (2012), Combining the Forecasts in the ECB SPF: Can anything beat the simple average?, forthcoming, *International Journal of Forecasting*

Geweke, J. and G. Amisano (2011), Optimal prediction pools, *Journal of Econometrics*, 164(1), 130-141

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(6), 1027-1043

Gneiting, T. F. Balabaoui and A. E. Raftery (2007), Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society*, 69(2), 243-268

Granger, C. W. J. and M. H. Peseran (2000), Economic and statistical measures of forecast accuracy, *Journal of Forecasting*, 19(7), 537-560

Harding, D. and A. Pagan (2010), Can we predict recessions?, Working Paper No. 69, *NCER Working Paper Series*, December.

Jore, A. S. J. Mitchell and S. P. Vahey (2010), Combining forecast densities from VARs with uncertain instabilities, *Journal of Applied Econometrics*, 25(2010), 621-634.

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

Kenny, G., T. Kostka and F. Masera (2011), How informative are the subjective expert forecasts of Macroeconomists?, *CESIFO Working Paper* No. 3671

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 (2012), How informative are central bank assessments of macroeconomic risks*, International Journal of Central Banking,* 8(3), 87-139.

Lahiri, K. and J. G. Wang (2007), The value of probability forecast as predictors of economic downturns, *Applied Economic Letters*, 14(14), 11-14

Mc Nees, S. K. (1975), An evaluation of economic forecasts. *New England Economic Review* (November/December), pp. 2-39.

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

Minzer J. and V. Zarnowitz (1969), The evaluation of economic forecasts, in Mincer, J. (ed), Economic Forecasts and Expectations, *National Bureau of Economic Research*, New York.

Murphy, A. H. (1973), A new vector partition of the probability score, *Journal of Applied Metreology*, 12, 595-600

Murphy, A. H. (1988), Skill scores based on the mean square error and their relationships to the correlation coefficient, *Monthly Weather Review*, 116, 2417–24.

Murphy, A. H. and R.L. Winkler (1992), Diagnostic verification of probability forecasts, *International Journal of Forecasting*, 7, 435–55.

Newey W. K. and K. D. West (1987), A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica*, 1987, vol. 55, issue 3, pages 703-08

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

Timmermann, A. (2006). Forecast combinations, Ch. 4. in G. Elliott, C.W.J. Granger and A. Timmermann (Eds.) Vol. 1, *Handbook of Economic Forecasting*, North-Holland.

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

Yates, J.F. (1982), Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 30, 132-156.

Figure 1: Outcome for target variable and events: GDP Growth

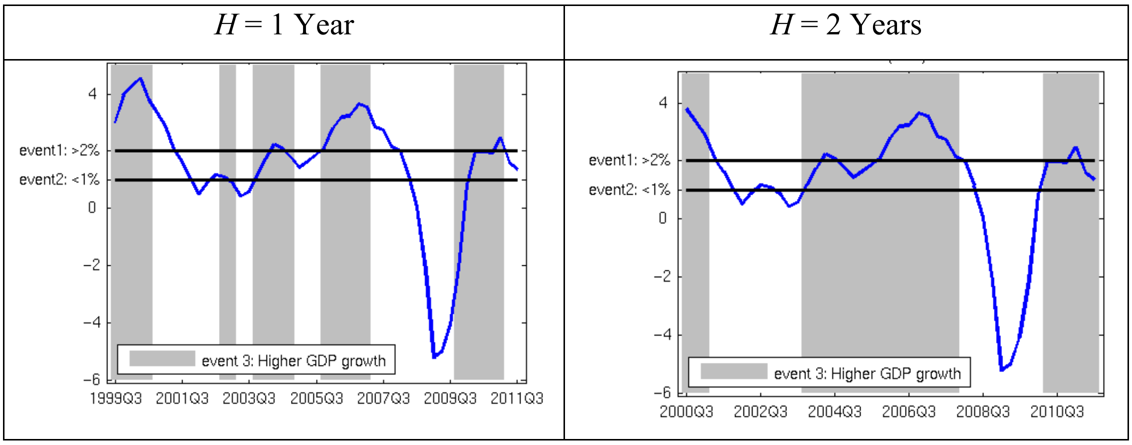


Figure 2: Outcome for target variable and events: HICP Inflation

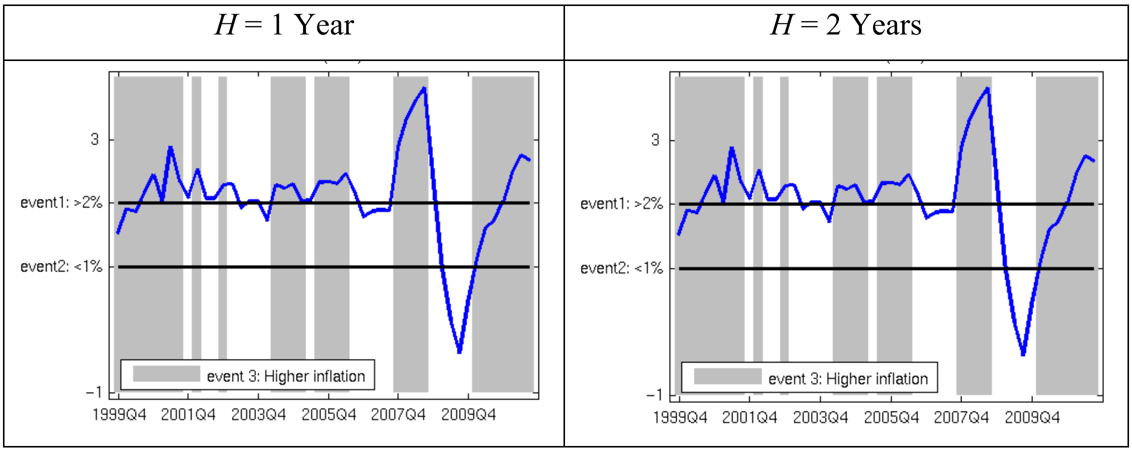
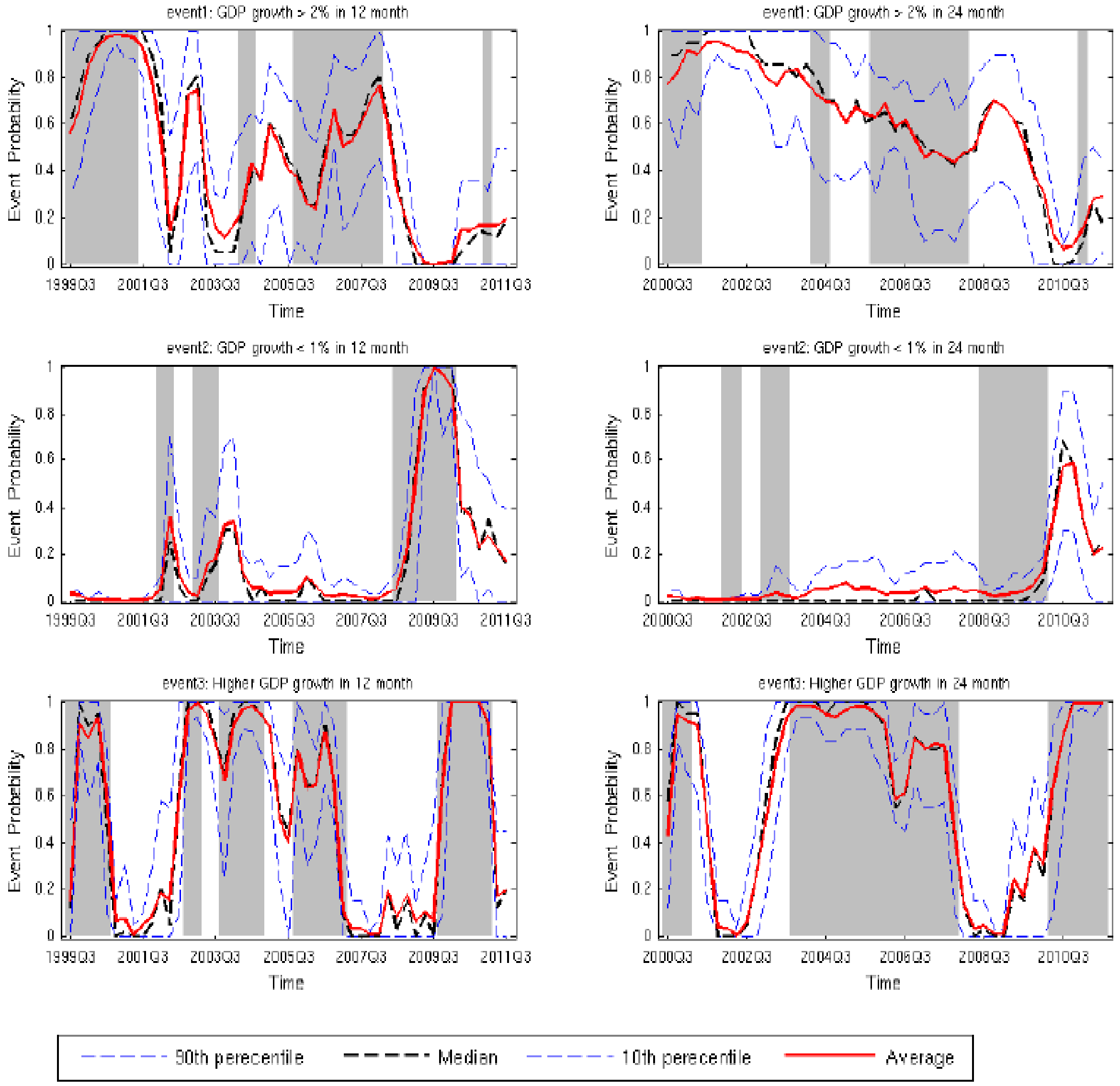
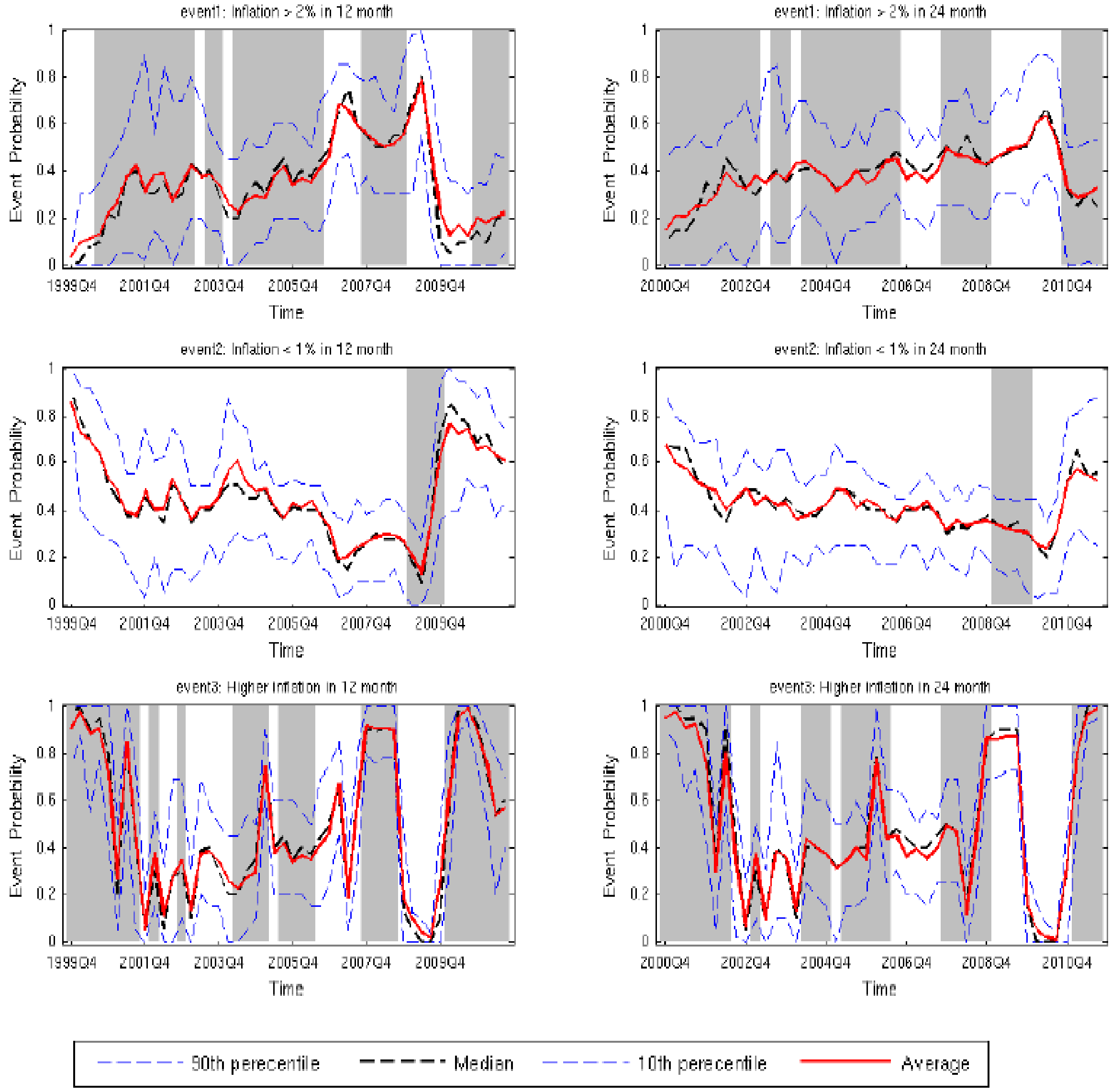


Figure 3: Probability forecasts (Median, 10 and 90% percentiles) for GDP events



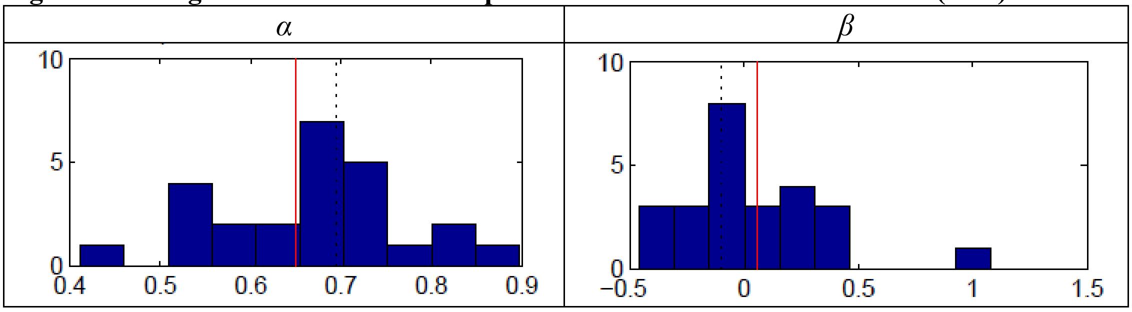
Note: The shaded region indicates the periods in which the event related to each corresponding probability forecast actually occurred.

Figure 4: Probability forecasts (Median, 10 and 90% percentiles) for Inflation events



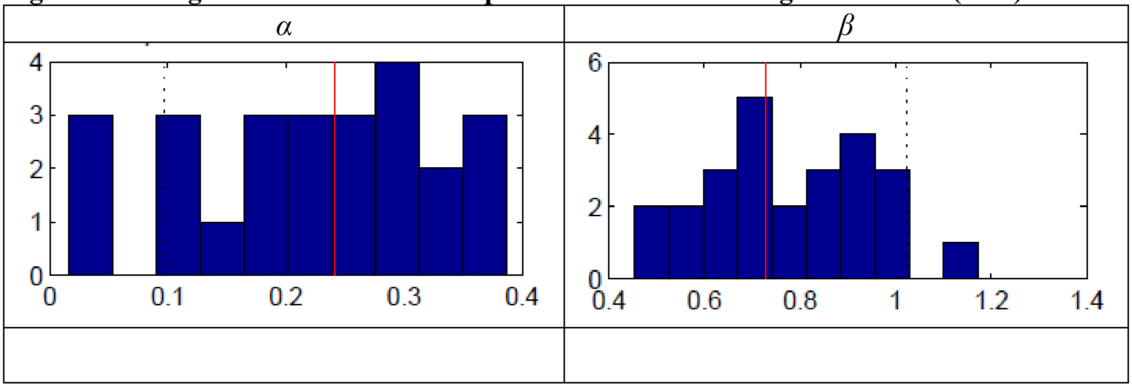
Note: The shaded region indicates the periods in which the event related to each corresponding probability forecast actually occurred.

Figure 5: Histogram of individual level parameter estimates: Inflation > 2% (H=1)



Note: The bars denote the number of forecasters for which the estimated parameter takes the value given on the horizontal axis. The vertical line depicts the estimated parameters based on the pooled regressions. The dotted line denotes the estimated parameters based on the probabilities from the aggregate distributions.

Figure 6: Histogram of individual level parameter estimates: Higher inflation (H=1)



Note: The bars denote the number of forecasters (indicated on the vertical axis) for which the estimated parameter takes the value given on the horizontal axis. The solid vertical line depicts the estimated parameters based on the pooled regressions. The dotted line denotes the estimated parameters based on the probabilities from the aggregate distributions.

Table 1: Decomposition of Quadratic Probability Score: Aggregate Growth Densities

Event	QPS =	Uncertainty +	Calibration Error	- Resolution
	$E[X-f]^2$	σ_x^2	$E_f\,[\mu_{x\mid f} - f]^2$	$E_f\,[\mu_{x\mid f} - \mu_x]^2$
			H = 1	
GDP growth > 2%	**0.38**	**0.49**	**0.01**	**0.12**
GDP growth < 1%	**0.25**	**0.36**	**0.01**	**0.12**
Higher GDP growth	**0.23**	**0.50**	**0.00**	**0.27**
			H = 2	
GDP growth > 2%	**0.65**	**0.45**	**0.21**	**0.00**
GDP growth < 1%	**0.49**	**0.36**	**0.15**	**0.02**
Higher GDP growth	**0.34**	**0.50**	**0.04**	**0.20**

Table 2: Tests of GDP Growth Events: Aggregate densities $x_{t+\tau} = \alpha + \beta f_{i,t+\tau} + \varepsilon_{i,t+\tau}$

	α	β	T	H$_0$: α =0, β =1	H$_0$: β =0
Forecast Horizon (H=1)					
GDP growth > 2%	*0.18* *(0.15)*	*0.56* *(0.27)*	*49*	*0.273*	*0.043*
GDP growth < 1%	*0.08* *(0.08)*	*0.93* *(0.25)*	*49*	*0.610*	*0.000*
Higher GDP growth	*0.04* *(0.10)*	*0.89* *(0.15)*	*49*	*0.744*	*0.000*
Forecast Horizon (H=2)					
GDP growth > 2%	0.44 (0.31)	-0.09 (0.47)	*45*	0.012	0.840
GDP growth < 1%	0.33 (0.12)	-0.91 (0.62)	*45*	0.002	0.150
Higher GDP growth	0.03 (0.13)	0.91 (0.16)	*45*	0.818	0.000

Note: Estimates of $x_{t+\tau}$= α+ β$f_{t+\tau}$ + $\varepsilon_{t+\tau}$ using Feasible Generalised Least Squares where $f_{t+\tau}$ denotes the probability forecasts extracted from the equal weighted aggregate SPF density. Standard errors corrected for serial correlation and aggregate shocks are reported in ( ).

Table 3: Tests of GDP Growth Events: Pooled individual densities

	α	β	N*T	H$_0$: α =0, β =1	H$_0$: β =0
Forecast Horizon (H=1)					
GDP growth > 2%	*0.29* *(0.03)*	*0.30* *(0.04)*	*1,071*	*0.000*	*0.000*
GDP growth < 1%	*0.13* *(0.02)*	*0.59* *(0.04)*	*1,071*	*0.000*	*0.000*
Higher GDP growth	*0.11* *(0.02)*	*0.74* *(0.03)*	*1,071*	*0.000*	*0.000*
Forecast Horizon (H=2)					
GDP growth > 2%	0.37 (0.04)	0.02 (0.06)	*927*	0.000	0.751
GDP growth < 1%	0.28 (0.02)	-0.47 (0.08)	*927*	0.000	0.000
Higher GDP growth	0.15 (0.02)	0.74 (0.03)	*927*	0.000	0.000

Note: Estimates of $x_{t+\tau}$= α+ β$f_{i,t+\tau}$ + $\varepsilon_{i,t+\tau}$ using Feasible Generalised Least Squares where $f_{i,t+\tau}$ denotes the individual level probability forecasts. Standard errors corrected for serial correlation and aggregate shocks are reported in ( ).

Table 4: Decomposition of Quadratic Probability Score: Aggregate Inflation Densities

Event	QPS =	Uncertainty +	Calibration Error	- Resolution
	$E[X\text{-}f]^2$	σ_x^2	$E_f\,[\mu_{x\|f} - f]^2$	$E_f\,[\mu_{x\|f} - \mu_x]^2$
			H = 1	
Inflation > 2%	**0.66**	**0.48**	**0.18**	**0.00**
Inflation < 1%	**0.46**	**0.24**	**0.23**	**0.01**
Higher inflation	**0.33**	**0.50**	**0.01**	**0.18**
			H = 2	
Inflation > 2%	**0.64**	**0.49**	**0.20**	**0.05**
Inflation < 1%	**0.48**	**0.21**	**0.34**	**0.06**
Higher inflation	**0.54**	**0.49**	**0.07**	**0.03**

Table 5: Tests of inflation events: Aggregate Density

	α	β	T	H$_0$: α =0, β =1	H$_0$: β =0
Forecast Horizon (H=1)					
Inflation > 2%	*0.73* *(0.25)*	*-0.19* *(0.63)*	*48*	*0.003*	*0.763*
Inflation < 1%	*0.27* *(0.21)*	*-0.40* *(0.34)*	*48*	*0.000*	*0.241*
Higher inflation	*0.11* *(0.12)*	*1.00* *(0.13)*	*48*	*0.254*	*0.000*
Forecast Horizon (H=2)					
Inflation > 2%	1.62 (0.28)	-2.38 (0.75)	*44*	0.000	0.004
Inflation < 1%	0.55 (0.29)	-1.10 (0.59)	*44*	0.000	0.068
Higher inflation	0.17 (0.13)	0.76 (0.26)	*44*	0.387	0.006

Note: Estimates of $x_{t+\tau} = \alpha + \beta f_{t+\tau} + \varepsilon_{t+\tau}$ using Feasible Generalised Least Squares where $f_{t+\tau}$ denotes the probability forecasts extracted from the equal weighted aggregate SPF density. Standard errors corrected for serial correlation and aggregate shocks are reported in ( ).

Table 6: Tests of inflation events: Pooled Individual Densities

	α	β	N*T	H$_0$: α =0, β =1	H$_0$: β =0
Forecast Horizon (H=1)					
Inflation > 2%	*0.65* *(0.03)*	*0.06* *(0.06)*	*1,028*	*0.000*	*0.284*
Inflation < 1%	*0.13* *(0.02)*	*-0.10* *(0.03)*	*1,028*	*0.000*	*0.002*
Higher inflation	*0.24* *(0.03)*	*0.73* *(0.04)*	*1,028*	*0.000*	*0.000*
Forecast Horizon (H=2)					
Inflation > 2%	0.85 (0.03)	-0.33 (0.07)	*895*	0.000	0.000
Inflation < 1%	0.13 (0.02)	-0.09 (0.05)	*895*	0.000	0.044
Higher inflation	0.26 (0.03)	0.59 (0.04)	*895*	0.000	0.000

Note: Estimates of $x_{t+\tau} = \alpha + \beta f_{i,t+\tau} + \varepsilon_{i,t+\tau}$ using Feasible Generalised Least Squares where $f_{i,t+\tau}$ denotes the individual level probability forecasts. Standard errors corrected for serial correlation and aggregate shocks are reported in ( ).

Table 7: Individual level tests of risk forecasts
(% of individuals for which hypothesis is rejected)

	H_0: $\alpha = 0, \beta = 1$		H_0: $\beta \leq 0$	
	H=1	**H=2**	**H=1**	**H=2**
GDP growth > 2%	*81%*	100%	*62%*	0%
	(81%)	*(100%)*	*(38%)*	*(0%)*
GDP growth < 1%	*54%*	100%	*96%*	0%
	(23%)	*(100%)*	*(96%)*	*(0%)*
Higher GDP growth	*38%*	36%	*100%*	100%
	(8%)	*(0%)*	*(100%)*	*(100%)*
Inflation > 2%	*100%*	100%	*8%*	4%
	(100%)	*(100%)*	*(0%)*	*(0%)*
Inflation < 1%	*100%*	100%	*0%*	0%
	(100%)	*(100%)*	*(0%)*	*(0%)*
Higher inflation	*40%*	50%	*100%*	96%
	(28%)	*(13%)*	*(100%)*	*(96%)*

Note: The table reports the number of individuals for which the hypotheses are rejected (at the α=10% level) expressed as a share of the total number of individuals in the panel. The test for zero resolution is based on a one sided t-test. The numbers in parentheses refers to the share of rejections after correcting for the false discovery rate in sequential hypothesis testing proposed by Benjamini and Hochberg (1995). In this set up, the Null is rejected at the α^{cor} = 10% * *i/N* level, *i* being the individual with the *i*-th lowest p-value from the set of individual level regressions.