Can Macroeconomists Forecast Risk? Event-Based Evidence from the Euro-Area SPF*

Geoff Kenny,[a] Thomas Kostka,[a] and Federico Masera[b]
[a]European Central Bank
[b]Universidad Carlos III de Madrid

We apply 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 more specific high and low 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 high and low outcome events, the surveys are most informative about GDP growth outcomes and at short horizons. The upper and lower regions of the predictive densities for inflation appear less informative.

JEL Codes: C22, C53.

*The authors would like to thank Domenico Giannone, Kajal Lahiri, Tom Stark, Simon van Norden, Shaun Vahey, Jonathan Wright, and Ken Wallis for useful comments on earlier drafts of this paper as well as participants at seminars in the Federal Reserve Bank of Philadelphia (June 29, 2012), the 32nd International Symposium on Forecasting (Boston, June 24–28, 2012), the ECB Workshop on SPF-related research on (November 26, 2012), the joint JCER-ESRI International Conference (Tokyo, February 21, 2013), and the 9th Annual joint Cirano-Cireq Workshop on data revisions in economic forecasting and policy (Montreal, October 10–11, 2013). We are also particularly grateful for many helpful comments and suggestions from three anonymous referees. 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. Corresponding author (Kenny): DG Research, European Central Bank, Kaiserstrasse 29, D-60311 Frankfurt, Germany. E-mail: Geoff.kenny@ecb.europa.eu.

1. Introduction

Recent experience with macroeconomic forecasting in an environment characterized 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 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. The Federal Reserve Bank of Philadelphia has an even longer tradition of collecting information on macroeconomists' assessments of future risks via its SPF. Indeed, a large literature has developed around the density forecasts from the U.S. SPF—see, for example, Diebold, Tay and Wallis (1999) Clements (2006, 2010), Giordani and Söderlind (2006), Lahiri and Wang (2007), and Engelberg, Manski, and Williams (2009), among others. In broad terms, these studies highlight the usefulness of the information contained in expert density forecasts—particularly at short horizons—whilst at the same time identifying notable shortcomings, including inattentiveness and overconfidence of forecasters when updating and reporting their assessments.

In this study we adopt a perspective similar to Clements (2006) and examine whether the SPF density forecasts from the ECB SPF provide insights about key future macroeconomic events. To do so, we partition the density at a given threshold point and consider only the binary set of mutually exclusive outcomes, i.e., either the outcome is above the specified threshold or it is below it. We focus on three key broad macroeconomic events: the probability of (i) a relatively low outcome of below 1 percent for the target variables (growth and inflation), (ii) a relatively high outturn of above 2 percent, and finally, (iii) an increase in the forecast target variable compared with its current level. By focusing on expert risk assessments of these "events," our analysis can reveal aspects of the densities

which are informative *even if* the overall density forecast exhibits a weak performance. This might arise, for example, as a result of heterogeneity in forecasters' loss functions which may lead them to be particularly adept (or inadept) at providing information on the likelihood of particular events. In helping identify aspects of the density forecasts which may be most reliable, our analysis is therefore of primary interest to density forecast users, including both monetary policymakers and those charged with maintaining financial stability.

Our evaluation is based on two key features of the densities. The first 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 is the main determinant of the densities' usefulness for decision making, as it measures their ability to discriminate between times when the risk materializes and times when it does not.[1] Our analysis is conducted using the aggregate headline SPF density forecasts that are regularly communicated to the public. However, we also pool density forecasts at the individual level and test for the presence of heterogeneity in forecaster skill at the individual level. To do so, we apply a GMM estimation procedure that allows for possible heterogeneity across forecasters as well as aggregate shocks impacting 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. In section 3, we provide some background descriptive statistics on the events and SPF probability forecasts. Section 4 presents our main empirical findings, while section 5 concludes.

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

[1] These two features are based on the decomposition of the quadratic probability score due to Murphy (1973). In an economic context, the Murphy decomposition has been used to evaluate probabilistic forecasts by Diebold and Rudebusch (1989), Lahiri and Wang (2007), and Galbraith and van Norden (2012) 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.

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 (1984), a long tradition exists of 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. They have been applied to economic forecasting by, among others, Berkowitz (2001), Clements (2006), Lahiri and Wang (2007), Boero, Smith, and Wallis (2011), Mitchell and Wallis (2011), and Galbraith and van Norden (2012) and are closely related to the field of interval forecasting discussed in Christoffersen (1998).[2] The approach we adopt here is also very much in the spirit of Berkowitz (2001) insofar as we emphasize 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 due to 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 $\text{QPS}(f_{t+\tau}, x_{t+\tau}) = E[f_{t+\tau} - x_{t+\tau}]^2$ thus provides a scoring rule which penalizes probability forecasts ($f_{t+\tau}$) which are low (high) when the event occurs (does not occur). To shed light on the attributes and validity of probability forecasts, Murphy (1973) suggested a factorization of the QPS based on the conditional distribution of $x_{t+\tau}$ given $f_{t+\tau}$, i.e.,

$$QPS(f_{t+\tau}, x_{t+\tau}) = \sigma_x^2 + E_f[\mu_{x/f} - f]^2 - E_f[\mu_{x/f} - \mu_x]^2. \quad (1)$$

The first term on the right-hand side of (1) measures the unconditional variance of the binary outcome variable, which can be seen

[2]Granger and Pesaran (2000) argue in favor 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.

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 the expected frequency of occurrence given the forecasts ($\mu_{x|f}$). Well-calibrated probability forecasts are approximately valid or "unbiased in the large" (Murphy and Epstein 1967). All other things equal, miscalibrated 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. The art of probability forecasting can thus be viewed as trying to minimize (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 miscalibration) that high-resolution forecasts may ultimately introduce.[3] 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. In the remainder of this section, we describe how we draw inference on these key properties of the SPF probabilities.

2.1 Aggregate-Level Analysis

Murphy and Winkler (1992), Galbraith and van Norden (2012), and Lahiri and Wang (2006, 2007) discuss econometric regression based tests of "perfect" calibration and "zero" resolution (i.e., no skill of forecasters in sorting outcomes). Such tests are based on a generalization of the forecast-realization regressions originally suggested in Mincer and Zarnowitz (1969) to probability forecasts. At

[3]See Lahiri and Wang (2007) and Galbraith and van Norden (2012) for further discussion of the Murphy decomposition.

the aggregate level, 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 for the same period:

$$x_{t+\tau} = \alpha + \beta f_{t+\tau} + \varepsilon_{i,t+\tau}. \tag{2}$$

As shown by Murphy and Winkler (1992), the fitted values of (2) provide an estimate of the expected probability of the outcome conditional on observing the forecast, i.e., $\mu_{x|f}$ in the decomposition of the QPS in (1). The estimate of forecast calibration, $E_f[\mu_{x|f} - f]^2$, is then computed as the sample averaged (i.e., mean) squared difference between the fitted value of this regression and the probability forecast. Likewise, an estimate of the resolution ($E_f[\mu_{x|f} - \mu_x]^2$) component in (1) corresponds to the mean squared difference between $\mu_{x|f}$ and an estimate for the unconditional mean of the outcome variable of the binary event (μ_x).[4]

As discussed in Murphy and Winkler (1992) and Galbraith and van Norden (2012), the hypothesis of perfectly calibrated forecasts requires that the forecasts do not systematically deviate from $\mu_{x|f}$, which is implied by the restrictions $\alpha = 0$ and $\beta = 1$. As it implies a joint restriction on the model's two parameters, the test of perfect calibration can be implemented as a Wald test using the χ^2 distribution with two degrees of freedom. As discussed in Holden and Peel (1990), albeit in a context that related to point forecasts, this test constitutes a sufficient condition for well-calibrated forecasts that also implies a significant degree of resolution; that is, under the null hypothesis ($\alpha = 0$ and $\beta = 1$), the outcome variable is positively correlated with the forecast. Moreover, joint hypotheses tests of this type may suffer from low power, especially in small samples. For these reasons, Holden and Peel (1990) suggest a necessary and therefore less demanding condition for calibration, namely the direct analysis of the "forecast error." Following this logic, we

[4]In estimating equation (1) we consider the simple discretized version originally suggested in Murphy and Winkler (1992) and Galbraith and van Norden (2012). In contrast, Galbraith and van Norden (2011) apply a non-parametric kernel regression of x on f. The advantage of such a continuous function is that it allows evaluation of calibration at any point in the continuous interval [0, 1]. In the present case where the forecast distribution is condensed to only two outcome bins, the two regressions would render the exact same results. Hence, a single regression of the binary outcome variable suffices in our setting.

also compute an additional test of calibration and test directly for a zero mean in the probability forecast error $(x_{t+\tau} - f_{t+\tau})$. This corresponds to a t-type test of $\alpha = 0$, conditional on the assumption that $\beta = 1(\alpha = 0|\beta = 1)$. To provide an intuition for the two alternative tests, a constant forecast that coincides with the unconditional outcome mean (μ_x) and therefore exhibits a zero mean forecast error would be considered well calibrated, or rather *unbiased*, according to the direct test of a zero mean of the probability forecast error $(x_{t+\tau} - f_{t+\tau})$. However, one would reject the null hypothesis $\alpha = 0$ and $\beta = 1$, as the constant forecast exhibits no correlation with the outcome variable. To avoid confusion of terminology, we refer to the sufficient condition ($\alpha = 0$ and $\beta = 1$) as perfect calibration and to the necessary condition $(\alpha = 0|\beta = 1)$ as a test of unbiasedness.

While the hypothesis of a perfectly calibrated forecast implies some positive correlation between the outcome's occurrence and the associated probability forecast, it is also of interest to independently test directly for such a correlation in its own right. Galbraith and van Norden (2012) point out that zero resolution, as defined in (1), requires $\beta = 0$. Under this restriction, the outcome's occurrence and its probability forecast are uncorrelated and, hence, the forecasts do not provide any signal on the event's likelihood of occurrence, i.e., $\mu_{x|f} = \mu_x$. Importantly, this test of zero resolution does not distinguish whether the outcome and the forecast probabilities are positively or negatively correlated. Clearly, however, it is of interest to examine the sign of the estimated β with good forecasts also exhibiting a positive correlation with the outcome. Hence, we focus on a simple one-sided t-test against the null hypothesis of "zero or negative signaling power" or $\beta \leq 0$. Rejection of this null suggests that the forecasts are useful, as they combine the idea of positive resolution with a non-negative correlation between the forecast and the event's occurrence. This latter one-sided test has the additional advantage that in small samples it is necessarily more powerful compared with the two-sided test of zero resolution ($\beta = 0$) studied in Galbraith and van Norden (2012).

Ideally, for good probability forecasts, we would want to *accept* both the tests of perfect calibration but *reject* the restrictions implied by zero or negative signaling power. In testing these hypotheses with the aggregate data, we estimate (2) by OLS whilst controlling for the serial correlation induced by the multi-period nature of

the forecast by using the correction to the standard errors of the parameters suggested by Newey and West (1987). However, Pesaran and Timmermann (2009) have recently highlighted that tests of the significance of ρ_{fx} (the correlation between x and f) are potentially distorted due to the clustering and serial correlation in the observed values of x. They propose a corrected test based on a dynamically augmented reduced rank regression which Lahiri and Wang (2013) have recently applied in examining the resolution of U.S. SPF forecasts. Hence, as an additional robustness check, we supplement the basic version of the one-sided test with the correction suggested by Pesaran and Timmerman (2009). The correction is implemented by a regression of $(\rho_{fx}/\sigma_x^2)x$ on f, augmented by the lagged values of the dependent and the independent variable.[5] As these two additional regressors contain information about the lagged forecast error, the Pesaran and Timmermann (2009) specification tests for the usefulness of the probability forecast conditional on such lagged information.

2.2 Panel Analysis

In extending the analysis of equation (2) to individual-level data, it is necessary to consider a more general model. In particular, heterogeneity in individual forecaster skill could be associated with different intercept and slope parameters across forecasters. In addition, as highlighted in earlier studies of surveyed point forecasts by Davies and Lahiri (1995), Clements, Joutz, and Stekler (2007), and earlier in Keane and Runkle (1990), an additional complication arises with individual-level data due to the role of aggregate macroeconomic shocks which can result in strong co-movement in forecast errors across individuals. A failure to account for such correlation could bias the estimated coefficients and standard errors and hence render unreliable any inference on the attributes of the SPF probability forecasts. In controlling for these aspects, we employ a GMM procedure similar to that described in Davies and Lahiri (1995), which generalized the approach of Keane and Runkle (1990) to allow for individual forecaster heterogeneity. In particular, we estimate a benchmark panel model that is equivalent to (2) and pools the

[5]See footnote 11 in Lahiri and Wang (2013).

estimated coefficients across individuals but allows for a very flexible structure of the residuals' variance-covariance matrix. The panel version is given in (3) below.

$$x_{t+\tau} = \alpha_{+}\beta f_{i,t+\tau} + \varepsilon_{i,t+\tau} \tag{3}$$

Using X to denote the $NT \times 1$ matrix containing N stacked output variables, and F to denote the corresponding $NT \times 2$ matrix for the regression constant and the stacked individual probability forecasts, the vector of estimated regression parameters ($\hat{\beta}_{GMM} = [\alpha, \beta]$) and their variance-covariance matrix (V) are given by $\hat{\beta}_{GMM} = \left[F'Z(Z'\hat{\Omega}^{-1}Z)^{-1}Z'F\right]^{-1}\left[F'Z(Z'\hat{\Omega}^{-1}Z)^{-1}Z'X\right]$ and $Var[\hat{\beta}_{GMM}] = (F'Z)^{-1}(F'\hat{\Omega}^{-1}F)(Z'F)^{-1} = V$, where Z is the set of instruments (see Keane and Runkle 1990). In our implementation, we take $Z = F$, and the estimators are given by $\hat{\beta}_{GMM} = \left[F'F(F'\hat{\Omega}^{-1}F)^{-1}F'F\right]^{-1}\left[F'F(F'\hat{\Omega}^{-1}F)^{-1}F'X\right]$ and $Var[\hat{\beta}_{GMM}] = (F'F)^{-1}(F'\hat{\Omega}^{-1}F)(F'F)^{-1} = V$.

In deriving the above GMM estimators, an essential step is the derivation of a correct residual covariance matrix ($\hat{\Omega}$). In specifying the elements of $\hat{\Omega}$, we follow Davies and Lahiri (1995) and allow for the variance of the regression residuals to potentially differ across individuals, i.e., $Var(\hat{\varepsilon}_{i,t}) = \sigma_i$ for $i = 1, \ldots N$. This approach generalizes the earlier approach in Keane and Runkle (1990) which had assumed homogeneity of error variances across individuals. Within our pooled panel regression framework where the estimated slope and intercept parameters are assumed constant across individuals, such a homogeneity assumption is unattractive, as it implies that no forecaster is systematically better than any other forecaster. Although we have no prior information that would lead us to believe some forecasters are systematically better than others, this later assumption is better tested rather than assumed a priori. Appendix 1 outlines precisely the structure of the assumed variance-covariance matrix. Moreover, as described below in section 2.3, in reporting our results we provide an explicit test of the homogeneity of the estimated coefficients and error variances across individuals following procedures suggested by Hsiao (2003) and Davies and Lahiri (1995), respectively.

2.3 Specification Tests and Forecast Diagnostics in the Panel Regression

In this subsection, we describe some of the forecast diagnostic tests that we can apply to our panel regressions. A first important test is of our maintained hypothesis that the intercept and slope parameters of equation (3) do not vary across individuals. We perform this test of our basic model following Hsiao (2003). Allowing for heterogeneity in the forecast bias (α_i) and the slope parameter (β_i), we commence with an unrestricted representation of the model which is given by (4).

$$x_{t+\tau} = \alpha_i + \beta_i f_{i,t+\tau} + \varepsilon_{i,t+\tau} \tag{4}$$

In (4) each of the coefficients are estimated separately by OLS. A test of the validity of the pooled model, proposed by Hsiao (2003), is based on the difference of the sample sum of squared residuals (RSS) of the unrestricted model and the pooled model ($\alpha_i = \alpha$ and $\beta_i = \beta$ for $i = 1, \ldots, N$)). In particular, the test statistic (F) given in (5) below has a Fisher distribution with $2(N-1)$ and $(NT - 2N)$ degrees of freedom under the null hypothesis of homogenous intercepts and slopes (i.e., $\alpha_i = \alpha$ and $\beta_i = \beta$ for $i = 1, \ldots N$).

$$\begin{aligned} F = {} & (RSS^{pooled} - RSS^{unrestrcited}) \\ & \times \frac{(RSS^{pooled} - RSS^{unrestricted})/2(N-1)}{RSS^{unrestricted}/(NT-2N)} \end{aligned} \tag{5}$$

Rejecting the null hypothesis implies that the pooling assumption is not valid because heterogeneous parameters are required to explain the variability in the panel data set. Hence, this test provides direct empirical evidence on the validity of the pooled coefficients implicit in our benchmark model (3). In reporting our results below, we also explore directly the estimated heterogeneity in individual parameter estimates when relevant.

Conditional on the pooled model's assumptions, it is also possible to examine tests of perfect calibration and positive signaling power discussed previously. As in the aggregate-level analysis, the test of perfect calibration implies a joint restriction on the pooled model's two parameters under the null hypothesis and can be implemented as a Wald test using the χ^2 distribution with two degrees of freedom. The alternative, less restrictive test for unbiasedness is

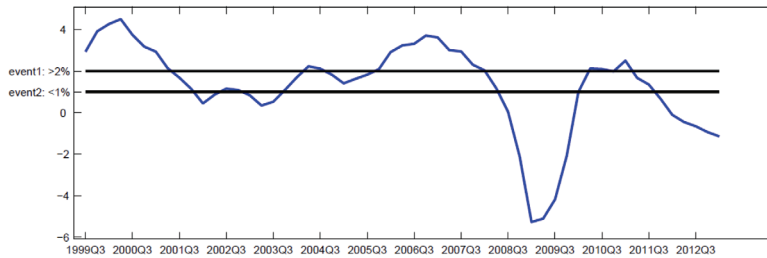
implemented as a t-test on the constant parameter (α) in the regression $x_{t+\tau} - f_{i,t+\tau} = \alpha_i + \varepsilon_{i,t+\tau}$. In the case of the test against the null hypothesis of zero or negative signaling power ($\beta \leq 0$), we can use a simple one-sided t-type test with $N^*T - 2$ degrees of freedom. In addition, given that our general model allows for individual-specific terms in the $\hat{\Omega}$ matrix, we can also test for the homogeneity of the error variances across individuals. Following Davies and Lahiri (1995), such a test can be performed by regressing $\hat{\varepsilon}_{i,t}^2$ on a constant and $N - 1$ individual dummies. The resulting R^2 multiplied by NT is distributed χ_{N-1}^2 under the null hypothesis of homoskedasticity with respect to the cross-sectional dimension, i.e., $Var(\hat{\varepsilon}_{i,t}) = \sigma_i = \sigma$ for all $i = 1, \ldots N$.

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 data set 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. The analysis is conducted using a filtered SPF data set which excludes *irregular* respondents as described in Genre et al. (2013) and draws on the quarterly rounds of the SPF conducted over the period 1999:Q1–2013:Q2. While such a fifteen-year period provides sufficient data to conduct our empirical tests, it can nonetheless only provide some first tentative signals on the quality of information that can be extracted from the survey. In particular, our results are subject to the significant caveat that they may be sensitive to the arrival of more data—a point emphasized in previous evaluations of the U.S. SPF conducted by Diebold, Tay, and Wallis (1999) and Croushore (2010).[6] As such, the data comprises

[6] Croushore (2010) finds in particular that the apparent bias found in earlier studies of the inflation forecasts in the U.S. SPF has dissipated as more data has become available. Similarly, the study by Diebold, Tay, and Wallis of the

Figure 1. Outcome for Target Variable: GDP Growth



a cross-sectional dimension of twenty-four to twenty-six 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 data set 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. As discussed in Genre et al. (2013), our focus on regular participants in the survey is motivated by the need to avoid sampling bias whereby individual forecaster performance might be excessively driven by a failure to reply during a particular business-cycle period. Our focus on regular respondents also ensures that the number of time-series observations for a given individual is maintained at a reasonably high level (an aspect that is important given the small-sample caveat highlighted above). More complete descriptions of the SPF data set, including a description of its panel dimension, are given in Bowles et al. (2010) and Genre et al. (2013). Garcia (2003) provides an earlier "bird's-eye" description of the ECB SPF.

Figures 1 and 2 provide time-series plots of the outcomes for the two forecast target variables, where horizontal lines are used to identify the lower and upper fixed thresholds of 1 percent and 2 percent,

U.S. SPF pointed to time variation in density forecast calibration error. In line with this, it cannot be excluded that some of our findings could be reversed by future researchers who can exploit sample sizes greater than the fifteen years that are currently available.

Figure 2. Outcome for Target Variable: HICP Inflation

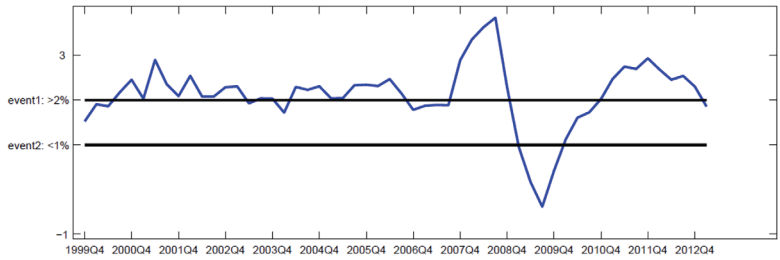


Table 1. GDP Growth Events and Probabilities: Summary Statistics

Event	T	Σx_t	Σx_t^*	μ_x	μ_f	$\rho_{x,f}$
			$H = 1$			
GDP Growth $> 2\%$	55	24	4	0.44	0.41	0.43
GDP Growth $< 1\%$	55	18	4	0.33	0.22	0.64
Higher GDP Growth	55	23	5	0.42	0.48	0.66
			$H = 2$			
GDP Growth $> 2\%$	51	20	4	0.39	0.58	-0.15
GDP Growth $< 1\%$	51	18	4	0.35	0.09	-0.07
Higher GDP Growth	51	26	3	0.51	0.59	0.74

respectively. To complement this graphical information, tables 1 and 2 provide average sample period information on the events and the forecast probabilities. In the case of GDP growth, for example, the chart indicates four occasions during which growth exceeded the 2 percent threshold we use for the analysis. This is confirmed also in table 1, where for this event we observe that x_t has taken a value of unity in twenty-four of the total of fifty-five outcomes included in our sample (i.e., $\Sigma x_t = 24$).[7] However, as discussed earlier, x_t

[7]While the SPF probabilities are fully real time in nature, our analysis conditions on the 2013:Q3 vintage for the HICP and GDP outcome series in defining our event outcome series x_t. An additional element of uncertainty for the evaluation relates to the possible impact of alternative data vintages for the occurrence and non-occurrence of the outcome. This may be most relevant for the GDP

Table 2. Inflation Events and Probabilities: Summary Statistics

Event	T	Σx_t	Σx_t^*	μ_x	μ_f	$\rho_{x,f}$
H = 1						
Inflation > 2%	54	37	6	0.69	0.36	−0.06
Inflation < 1%	54	4	1	0.07	0.09	−0.03
Higher Inflation	54	31	7	0.57	0.43	0.59
H = 2						
Inflation > 2%	50	36	5	0.72	0.39	−0.51
Inflation < 1%	50	4	1	0.08	0.07	−0.30
Higher Inflation	50	29	6	0.58	0.47	0.62

exhibits high persistence and switches from zero to unity on only four occasions, $\sum x_t^* = 4$ where $x_t^* = 1$ *if* $x_t = 1$ *and* $x_{t-1} = 0$, and 0 otherwise. Similarly, there were also four occasions when growth fell below the lower threshold of 1 percent, most notably during the Great Recession of 2008 and 2009. At the one-year horizon, the summary statistics in table 1 illustrate that 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.[8]

In the case of inflation, the pattern is somewhat different, with annual inflation being quite often above the 2 percent threshold we

events, while inflation in the euro area has been less subject to revision (see, for example, the recent study by Kenny, Kostka, and Masera 2014a). We have therefore computed many of the tests we report later in the paper using the first estimate to define the event outcome, and overall we have reached broadly similar conclusions to those that are reached based on the 2013:Q3 vintage.

[8]In the empirical analysis even for the fixed threshold events, the outcome's frequency of occurrence can change depending on the forecast horizon because the sample size changes. For the two-year-ahead forecast horizon, the sample size for the GDP forecasts is reduced to fifty-one compared with fifty-five for the one-year-ahead forecasts (for inflation the corresponding reduction is from fifty-four to fifty). In the case of the direction-of-change forecasts, the statistics on the outcome's occurrence/non-occurrence necessarily change with the horizon, which is a defining characteristic of the event. In other words, higher GDP growth/inflation in two years' time is a conceptually distinct event from higher GDP growth/inflation in one year's time.

use for this study. For example, as shown in table 2, in thirty-seven of the fifty-four periods studied, inflation was above 2 percent, i.e., 69 percent of the time when measured over the one-year horizon (i.e., $\mu_x = 0.69$). Given this outcome, if the probability assessments of SPF participants are well calibrated, we might expect to observe relatively high probabilities for this event. As shown in table 2, the average probability assigned by SPF respondents stood at 36 percent over the sample ($\mu_f = 0.36$), providing some first-pass evidence that they may have under-estimated this probability. In contrast, the below 1 percent outcome for inflation has occurred only once during the sample (during the 2009 downturn linked to the recent financial crisis), although this occurrence persisted for four quarters. For aggregate analysis, there are considerable limitations and caveats that apply when empirically testing the information content of expert probability assessments, given that we observe the event only once in the sample. In particular, such tests are likely to have relatively modest power and may be biased. However, given that we employ panel data capturing imperfectly correlated expert assessments, in the subsequent analysis we are able to conduct statistical inference concerning the ability of macroeconomists to assess the likelihood of even these relatively 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 observed at the time of the survey has also occurred quite frequently.

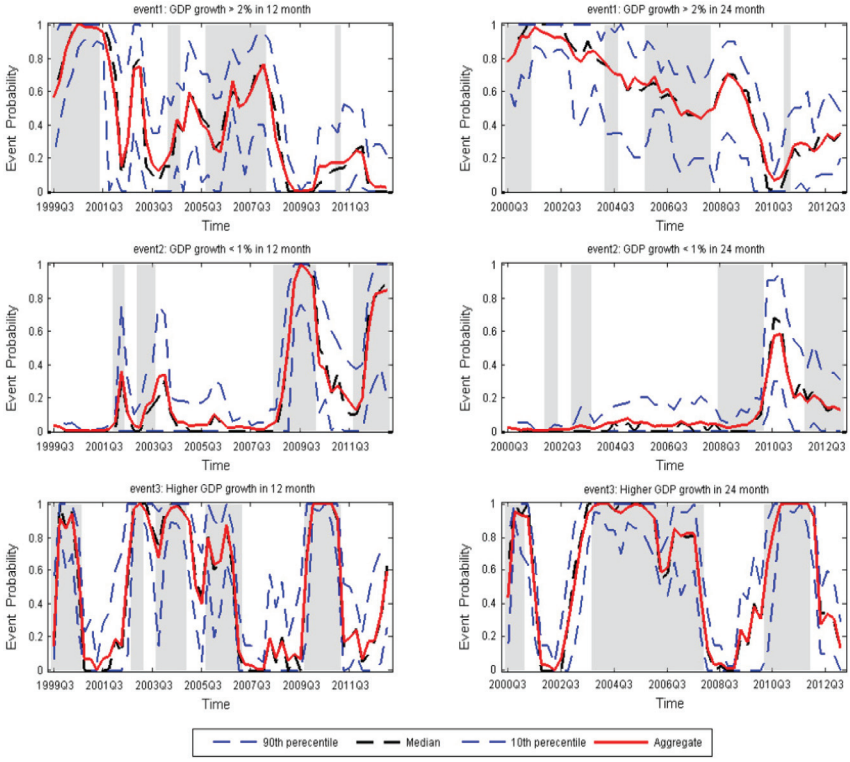
To conduct our analysis, we also need to extract the cumulative probabilities for the three events that we analyze. SPF respondents submit their replies in the form of discrete histograms assigning probabilities to a set of intervals representing possible outcome ranges for the target variable. In addition, 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.

Figures 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 periods during 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. Indeed, as also shown in tables 1 and 2, all of the direction-of-change forecasts exhibit a positive correlation with the outcome, i.e., ρ_{xf} is 0.66 and 0.74 for GDP H = 1 and H = 2, respectively, and 0.59 and 0.62 for the corresponding inflation forecasts. The positive correlation suggests that even at the longer horizons experts can accurately gauge the cyclical tendency of the economy compared with its current state. At the same time, these simple correlations need to be interpreted with due caution and warrant a more robust econometric evaluation, which we provide in the next section.

This first visual inspection of the probability forecasts in figures 3 and 4 suggests a less clear correspondence between the expert probability assessments for the high/low fixed threshold events. 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

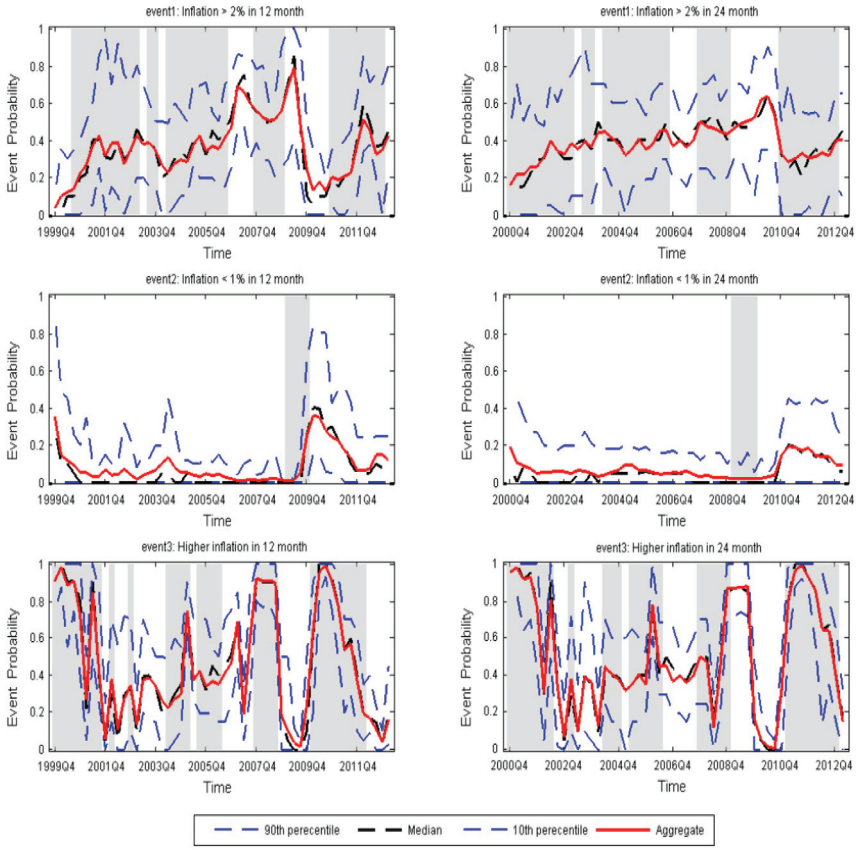
Figure 3. Probability Forecasts (Median, 10th, and 90th Percentiles) for GDP Events



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

completely passed. This graphical evidence of the weaker relationship for the fixed threshold events is also highlighted by a relatively small negative correlation between the aggregate probability forecasts and the observed outcome in table 2. The graphical evidence for the fixed threshold GDP events also suggests relatively limited and/or lagging signaling power of the SPF probability assessments. An exception is perhaps the one-year-ahead GDP predictions for both high and low outcomes which, as observed in table 1, also exhibit a positive correlation of 0.43 and 0.64 with the high and low outcomes, respectively.

Figure 4. Probability Forecasts (Median, 10th, and 90th Percentiles) for Inflation Events



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

4. Evaluation Results for SPF Event Forecasts

In this section, we exploit the QPS decomposition and calibration and resolution tests discussed previously in order to shed more robust econometric evidence on the information content of these attributes of the SPF densities. We first report the QPS decomposition, providing evidence of miscalibration and signaling power.

We then report more formal tests of perfect calibration and unbiasedness and tests of zero or negative signaling power using both aggregate and pooled individual-level data. Finally, we explore the heterogeneity in the SPF panel in more detail. We consider first the results for the GDP events and then subsequently turn to inflation.[9]

4.1 GDP Growth Events

Table 3 reports the QPS decomposition for each of the three GDP events. The decomposition is based on the aggregated probabilities as published on a quarterly frequency by the ECB. The QPS statistics indicate that the SPF densities perform less well at capturing the more extreme threshold events, whilst the direction-of-change predictions have lower QPS and hence performed 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 resolution, which is particularly strong for the direction-of-change forecasts. In contrast, the signaling information for the upper and lower fixed threshold events is smaller. At the two-year horizon, there is a notable increase in the miscalibration for the fixed threshold event forecasts, while the calibration error for the direction-of-change forecast continues to be very small. The latter forecasts also continue to possess useful signaling information, as reflected in the 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 specific high/low GDP growth events that lie further ahead in the future. 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 signaling power described in section 2.

Table 4 reports the results from the estimation of equation (2) based on the aggregate probabilities. 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

[9] Appendix 2 presents some sensitivity analysis where we vary the fixed thresholds used to define the high and low outcome events.

Table 3. Decomposition of Quadratic Probability Score: GDP Events

Event	QPS = $E[X-f]^2$	Uncertainty + σ_x^2	Calibration Error $E_f[\mu_{x\|f}-f]^2$	− Resolution $E_f[\mu_{x\|f} - \mu_x]^2$
H = 1				
GDP Growth > 2%	0.38	0.49	0.01	−0.12
GDP Growth < 1%	0.25	0.37	0.01	−0.13
Higher GDP Growth	0.24	0.50	0.00	−0.27
H = 2				
GDP Growth > 2%	0.71	0.47	0.24	0.00
GDP Growth < 1%	0.54	0.39	0.17	−0.02
Higher GDP Growth	0.20	0.49	0.00	−0.29

Table 4. Tests of GDP Growth Events: Aggregate Densities

				(1) H_0:	(2) H_0:	(3) H_0:	(4) H_0:
	T	α	β	$\alpha = 0, \beta = 1$	$\beta \leq 0$	$\alpha = 0 \vert \beta = 1$	$\gamma \leq 0$
Forecast Horizon (H = 1)							
GDP Growth > 2%	55	0.15	0.70	0.327	0.001	0.755	0.034
		(0.12)	(0.22)				
GDP Growth < 1%	55	0.11	1.02	0.244	0.000	0.106	0.001
		(0.07)	(0.10)				
Higher GDP Growth	55	0.01	0.85	0.542	0.000	0.425	0.000
		(0.07)	(0.14)				
Forecast Horizon (H = 2)							
GDP Growth > 2%	51	0.56	−0.29	0.000	0.787	0.212	0.013
		(0.25)	(0.37)				
GDP Growth < 1%	51	0.38	−0.25	0.038	0.663	0.016	0.095
		(0.15)	(0.61)				
Higher GDP Growth	51	−0.06	0.97	0.521	0.000	0.306	0.002
		(0.06)	(0.11)				

Notes: Estimates of α and β are obtained from the OLS regression of $x_{t+\tau} = \alpha + \beta^* f_{t+\tau} + \varepsilon_{t+\tau}$, where $f_{t+\tau}$ denotes the probability forecasts extracted from the equal weighted aggregate SPF density and $x_{t+\tau}$ the corresponding binary outcome indicator. Newey-West standard errors correcting for serial correlation are given in parentheses. Columns 1 and 2 report p-values for the indicated hypotheses. The p-values in column 3 are obtained from the OLS regression $x_{t+\tau} - f_{t+\tau} = \alpha + \varepsilon_{t+\tau}$. The p-values in column 4 are obtained from the OLS regression $\theta x_{t+\tau} = c(1-\phi) + \gamma^* f_{t+\tau} - \phi\gamma^* f_{t+\tau-1} + \phi\theta^* x_{t+\tau-1} + \varepsilon_{t+\tau}$, where $\theta = \rho_{x,f}/\sigma_x^2$.

(H_0: $\alpha = 0, \beta = 1$) 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 the values under the null hypothesis. Not surprisingly, therefore, the Wald test is unable to reject. For the three events considered, we can also reject the null hypothesis of negative or zero signaling power (H_0: $\beta \leq 0$) at H = 1. This tends to confirm the useful signaling information in the SPF densities for GDP at short horizons, both for the direction-of-change and the fixed threshold events. However, at longer horizons, the above findings are reversed. For both high and low fixed threshold events, we reject the null of perfect calibration at the 5 percent level. The null of zero or negative signaling power cannot be rejected for both the relatively high and low growth thresholds at the two-year horizon. The estimates of β for these events at this longer horizon also tend to be negative, implying that the probability forecasts tend to fall when these high/low GDP growth events occur and vice versa. 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 or negative signaling power. 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.

In general, the aggregate analysis discussed above points to a high degree of information content in SPF probability forecasts for the direction of change in GDP growth at both short and longer horizons. This finding also appears robust when one considers other variants of the tests discussed in section 2. For example, the direct test of the zero mean in the probability forecast error cannot be rejected, while a test based on the Pesaran and Timmermann (2009) correction for clustering in the outcome variable also rejects the null of zero or negative signaling power in the direction-of-change forecasts. Also, when examining the results of these other tests (reported in columns 3 and 4 in table 4), the evidence highlights the relatively good performance of the fixed threshold event forecasts at short horizons (H = 1). Indeed, in contrast to our baseline result, the Pesaran and Timmermann (2009) specification suggests that the hypothesis of zero or negative signaling power for the high/low GDP growth

event forecasts is rejected even at longer horizons. This result implies that the probability forecast contains some information beyond the information that is contained in the lagged outcome and the lagged forecast.[10]

Table 5 reports the equivalent regression results for all three GDP event forecasts but based on the unbalanced panel of individual responses and including a correction for both serial correlation and aggregate shocks (using the GMM estimation procedure described in section 2 and appendix 1). In columns 1 and 2 we report the results of the tests for the commonality of coefficients in the panel regression model and the homogeneity of error variances, respectively. In general, these tests support the validity of pooling coefficients for all events considered and at both horizons. Similarly, controlling for aggregate shocks, we are unable to reject the assumption that the error variances are equal across forecasters. We interpret these findings as supporting the idea that the heterogeneity in SPF forecasts for GDP cannot be seen as evidence of significant differences in forecast performance of individual forecasts when viewed from an ex post perspective. This result is somewhat in line with the findings in a recent study by D'Agostino, McQuinn, and Whelan (2012) in relation to the point forecasts in the U.S. SPF. In particular, they find limited evidence for the idea that the best forecasters are actually significantly better than others, though there is evidence that a relatively small group of forecasters perform very poorly. Our finding above is also in line with recent analysis of ECB SPF densities conducted in Kenny, Kostka, and Masera (2014b), who show that for GDP growth a significant improvement in aggregate density performance cannot be achieved by simply excluding those forecasters with a relatively poor track record.

Regarding the key attributes of the forecasts, the panel results in table 5 tend to confirm many of the findings observed in table 4 using the aggregate-level data. For short horizons, the hypothesis that the probability forecasts exhibit zero or negative signaling power is strongly rejected (column 4). At longer horizons, however,

[10] Given the small sample available for estimation with the aggregate results and the additional parameters that need to be estimated in the implementation of this test, we would think that some caution is warranted in interpreting this result.

Table 5. Tests of GDP Growth Events: Pooled Individual Densities

				(1) H_0:	(2) H_0:	(3) H_0:	(4) H_0:	(5) H_0:
	N^*T	α	β	$\alpha_i = \alpha, \beta_i = \beta$	$\sigma_i = \sigma$	$\alpha = 0, \beta = 1$	$\beta \leq 0$	$\alpha = 0 \vert \beta = 1$
Forecast Horizon (H = 1)								
GDP Growth > 2%	1,172	0.23	0.49	1.000	0.999	0.047	0.009	0.813
		(0.13)	(0.21)					
GDP Growth < 1%	1,172	0.15	0.80	0.983	0.976	0.136	0.000	0.132
		(0.08)	(0.17)					
Higher GDP Growth	1,172	0.07	0.73	1.000	0.987	0.101	0.000	0.428
		(0.10)	(0.14)					
Forecast Horizon (H = 2)								
GDP Growth > 2%	1,095	0.52	−0.20	0.999	0.998	0.000	0.797	0.247
		(0.16)	(0.24)					
GDP Growth < 1%	1,095	0.36	−0.21	1.000	1.000	0.001	0.686	0.023
		(0.10)	(0.44)					
Higher GDP Growth	1,095	0.03	0.84	1.000	0.988	0.386	0.000	0.393
		(0.11)	(0.15)					

Notes: Estimates of α and β are obtained from the pooled GMM regression $x_{t+\tau} = \alpha + \beta^* f_{i,t+\tau} + \varepsilon_{i,t+\tau}$, where $f_{i,t+\tau}$ denotes the probability forecasts extracted from the SPF participant's individual density and $x_{t+\tau}$ the corresponding binary outcome indicator. GMM standard errors correcting for serial correlation and cross-sectional serial correlation are given in parentheses. Columns 1 to 5 report p-values for the indicated null hypotheses. The p-values in column 1 are obtained from the Hsiao test of parameter heterogeneity, which is based on a comparison of the residual sum of squares between the OLS regressions of the constrained model $x_{t+\tau} = \alpha + \beta^* f_{i,t+\tau} + \varepsilon_{i,t+\tau}$ and the unconstrained model $x_{t+\tau} = \alpha_i + \beta_i{}^* f_{i,t+\tau} + \varepsilon_{i,t+\tau}$ whereas the unconstrained model corresponds to N individual-specific regressions. P-values in column 2 are obtained from the F-test of the regression $\varepsilon^2_{i,t+\tau} = \gamma_i + \upsilon_{i,t+\tau}, \varepsilon_{i,t+\tau}$ being the residuals from the constrained model. The p-values in column 5 are obtained from the pooled GMM regression $x_{t+\tau} - f_{i,t+\tau} = \alpha + \varepsilon_{t+\tau}$.

the probability forecast for relatively high and low growth outcomes exhibits no useful signaling power and is inversely correlated with the outcome, as highlighted by negatively estimated slope parameters for these events. Although the low growth event does not correspond to a recession in a classical sense, 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 compliments the earlier findings based on aggregate probabilities, is the inability to reject the perfect-calibration hypothesis at the individual level for fixed threshold events at the one-year forecast horizon but the strong tendency to reject perfect calibration at the longer two-year horizon (column 3). As was shown with the aggregate-level analysis, the direction-of-change forecasts for GDP growth appear well calibrated at both short and longer horizons, also when basing the analysis on the less demanding hypothesis of a zero mean in the forecast error (column 5). In contrast, for the fixed threshold events, our empirical results point to a quite dramatic deterioration in the information value of density forecasts for real output growth when the forecast horizon is extended. The probability assessments of experts included in the SPF panel thus provide insightful information on the general cyclical evolution of the economy, but at longer horizons they are less informative about the occurrence or non-occurrence of specific high or low GDP growth events.

In assessing the above findings on the information content of the probability forecasts extracted from the ECB SPF, it is interesting to compare them with previous studies using U.S. data. For example, in recent studies of the U.S. SPF, Lahiri and Wang (2006, 2013) study the usefulness of probability forecasts for GDP declines. They find that the shorter-horizon forecasts for GDP growth are more informative than longer-horizon ones—a finding which directly echoes our results for the euro area. At the same time, our results point to noticeable limitations for the probability forecasts for specific fixed threshold events at longer horizons. Lahiri and Wang (2006) highlight in particular the limitations of the probability forecasts for less frequently occurring events such as recessions and point to the role of behavioral factors that may lead forecasters to inappropriately anchor their expectations in difficult forecasting environments.

Similar limitations on the information content of the SPF density forecasts for U.S. output growth are also highlighted in Clements (2010), who finds that not all forecasters update the information in their histograms as new information arrives. Finally, a key aspect of our analysis has been the exploration of heterogeneity in the performance of the densities across individuals. For the GDP densities, our panel results point to the poolability of the information in the density forecasts and hence a relatively low degree of heterogeneity across forecasters' performance when viewed from an ex post evaluation perspective.[11] Other studies have found stronger evidence of significant forecaster heterogeneity. For example, studies such as Boero, Smith, and Wallis (2008), Engleberg, Manski, and Williams (2009), and previously Davies and Lahiri (1995) emphasized the significance of such heterogeneity in relation to the individual point forecasts.

4.2 Inflation Events

Table 6 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 favorably with the previous findings for GDP. In particular, the inflation scores are, as a rule, noticeably higher than the corresponding GDP scores. Moreover, SPF probability forecasts for both high and low fixed threshold inflation events show signs of miscalibration—even at the shorter horizons. Similarly, the probability forecasts for the specific fixed threshold events exhibit quite low signaling power, as reflected in low resolution. However, once again, the direction-of-change forecasts appear better calibrated and exhibit higher resolution. This is mainly at the one-year horizon, however, and the evidence that the direction-of-change forecasts for inflation provide useful signals (i.e., high resolution) at longer horizons is weaker.

Tables 7 and 8 report the econometric tests for perfect calibration and zero resolution for each of the three inflation events for the aggregate- and panel-level analysis, respectively. At the aggregate level (table 7), the results for inflation contrast with some of the

[11] We explore further the empirical evidence on forecaster heterogeneity in section 4.3 below.

Table 6. Decomposition of Quadratic Probability Score: Inflation Events

	QPS =	Uncertainty +	Calibration Error −	Resolution
Event	$E[X-f]^2$	σ_x^2	$E_f[\mu_{x\|f}-f]^2$	$E_f[\mu_{x\|f}-\mu_x]^2$
H = 1				
Inflation > 2%	0.72	0.44	0.27	0.00
Inflation < 1%	0.54	0.15	0.39	−0.01
Higher Inflation	0.32	0.48	0.02	−0.18
H = 2				
Inflation > 2%	0.73	0.42	0.42	−0.11
Inflation < 1%	0.44	0.17	0.29	−0.02
Higher Inflation	0.41	0.50	0.02	−0.10

Table 7. Tests of Inflation Events: Aggregate Density

				(1) H_0:	(2) H_0:	(3) H_0:	(4) H_0:
	T	α	β	$\alpha = 0, \beta = 1$	$\beta \leq 0$	$\alpha = 0 \mid \beta = 1$	$\gamma \leq 0$
Forecast Horizon ($H = 1$)							
Inflation $> 2\%$	54	0.75	-0.17	0.001	0.607	0.003	0.419
		(0.26)	(0.63)				
Inflation $< 1\%$	54	0.08	-0.07	0.000	0.599	0.806	0.087
		(0.08)	(0.30)				
Higher Inflation	54	0.14	1.01	0.109	0.000	0.038	0.000
		(0.09)	(0.13)				
Forecast Horizon ($H = 2$)							
Inflation $> 2\%$	50	1.67	-2.46	0.000	0.999	0.006	0.112
		(0.27)	(0.72)				
Inflation $< 1\%$	50	0.20	-1.67	0.000	0.932	0.925	0.389
		(0.14)	(1.12)				
Higher Inflation	50	0.09	1.06	0.040	0.000	0.041	0.000
		(0.08)	(0.09)				

Notes: Estimates of α and β are obtained from the OLS regression of $x_{t+\tau} = \alpha + \beta^* f_{t+\tau} + \varepsilon_{t+\tau}$, where $f_{t+\tau}$ denotes the probability forecasts extracted from the equal weighted aggregate SPF density and $x_{t+\tau}$ the corresponding binary outcome indicator. Columns 1 and 2 report p-values for the indicated hypotheses. The p-values in column 3 are obtained from the OLS regression $x_{t+\tau} - f_{t+\tau} = \alpha + \varepsilon_{t+\tau}$. Newey-West standard errors correcting for serial correlation are given in parentheses. The p-values in column 4 are obtained from the OLS regression $\theta x_{t+\tau} = c(1-\phi) + \gamma^* f_{t+\tau} - \phi\gamma^* f_{t+\tau-1} + \phi\theta^* x_{t+\tau-1} + \varepsilon_{t+\tau}$, where $\theta = \rho_{x,f}/\sigma_x^2$.

Table 8. Tests of Inflation Events: Pooled Individual Densities

				(1) H_0:	(2) H_0:	(3) H_0:	(4) H_0:	(5) H_0:
	N^*T	α	β	$\alpha_i = \alpha, \beta_i = \beta$	$\sigma_i = \sigma$	$\alpha = 0, \beta = 1$	$\beta \leq 0$	$\alpha = 0 \vert \beta = 1$
Forecast Horizon (H = 1)								
Inflation > 2%	1,166	0.67	0.05	0.994	1.000	0.000	0.410	0.034
		(0.12)	(0.21)					
Inflation < 1%	1,166	0.07	0.09	0.776	1.000	0.000	0.353	0.844
		(0.06)	(0.23)					
Higher Inflation	1,166	0.24	0.77	0.639	0.722	0.018	0.000	0.058
		(0.09)	(0.17)					
Forecast Horizon (H = 2)								
Inflation > 2%	1,097	0.90	−0.43	0.001	1.000	0.000	0.999	0.047
		(0.08)	(0.14)					
Inflation < 1%	1,097	0.10	−0.36	0.921	1.000	0.000	0.914	0.825
		(0.06)	(0.26)					
Higher Inflation	1,097	0.19	0.82	0.924	0.930	0.004	0.000	0.004
		(0.08)	(0.19)					

Notes: Estimates of α and β are obtained from the pooled GMM regression $x_{t+\tau} = \alpha + \beta^* f_{i,t+\tau} + \varepsilon_{i,t+\tau}$, where $f_{i,t+\tau}$ denotes the probability forecasts extracted from the SPF participant's individual density and $x_{t+\tau}$ the corresponding binary outcome indicator. GMM standard errors correcting for serial correlation and cross-sectional serial correlation are given in parentheses. Columns 1 to 5 report p-values for the indicated null hypotheses. The p-values in column 1 are obtained from the Hsiao test of parameter heterogeneity, which is based on a comparison of the residual sum of squares between the OLS regressions of the constrained model $x_{t+\tau} = \alpha + \beta^* f_{i,t+\tau} + \varepsilon_{i,t+\tau}$ and the unconstrained model $x_{t+\tau} = \alpha_i + \beta_i^* f_{i,t+\tau} + \varepsilon_{i,t+\tau}$ whereas the unconstrained model corresponds to N individual-specific regressions. P-values in column 2 are obtained from the F-test of the regression $\varepsilon^2_{i,t+\tau} = \gamma_i + \upsilon_{i,t+\tau}$, $\varepsilon_{i,t+\tau}$ being the residuals from the constrained model. The p-values in column 5 are obtained from the pooled GMM regression $x_{t+\tau} - f_{i,t+\tau} = \alpha + \varepsilon_{t+\tau}$.

previous findings observed for GDP. In particular, in column 1, the test of well-calibrated forecasts tends to be rejected for the fixed threshold events at both horizons.[12] Also, we are unable to reject the null of negative signaling power (column 3) for all fixed threshold events. Moreover, as indicated by the estimated β parameters and the tests of negative signaling power (column 2, table 7), the SPF predictions for the fixed threshold 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 specific events toward the upper and lower range of inflation outcomes. Turning to the direction-of-change forecasts for inflation, the aggregate results tend to mirror some of the more positive findings that we observed for GDP. In particular, direction-of-change forecasts appear quite informative, as they are better calibrated and exhibit significant positive resolution. Even though the hypothesis of perfect calibration is often rejected at the 10 percent confidence level, the unrestricted estimates of α and β for the direction-of-change forecasts in table 7 correspond closely with their predicted values under the null.

The above aggregate results on the poor calibration of fixed threshold events for inflation are confirmed by the panel analysis in table 8 where the test of perfect calibration (column 3) is rejected for all four fixed threshold events and at both horizons. This finding for inflation compares less favorably with the equivalent analysis for GDP, where the fixed threshold probabilities were shown to be better calibrated at short horizons (H = 1). Moreover, and in line with the aggregate-level analysis, as indicated by the estimated β parameters and the tests of negative signaling power (column 4, table 8), the SPF probabilities for the fixed threshold inflation events appear to be either uncorrelated or correlate negatively with the occurrence of these events. Also mirroring the aggregate-level analysis,

[12]For the fixed threshold inflation events, the direct test of the mean probability error in column 3 of table 8 provides a conflicting signal to the test in column 1 of table 8. This discrepancy reflects the large deviation of the unrestricted estimates of β from its value of unity under the null in the direct test. Given the evidence of noticeable miscalibration in table 6, we would tend to trust more the results of the test in column 1.

direction-of-change forecasts appear quite informative, as they are better calibrated and exhibit significant positive resolution. For both forecast horizons, the test of negative signaling power can be rejected in contrast to the equivalent fixed threshold events. Once again, this finding highlights the ability of SPF panel members to capture the broad cyclical direction of inflation but a much weaker ability to correctly diagnose the likelihood of more specific high and low inflation outcomes. Concerning the heterogeneity of the SPF forecasts for inflation, the test results reported in table 8 suggest that the pooling of individual coefficients and error variances is, for most events, not rejected by the data. One interesting exception is the event for high inflation outcomes above 2 percent where the common slope and intercept in the panel model is rejected (although the common error variance is not). These differences in individual forecaster attributes for this event, which is of some interest from a monetary policy perspective, are discussed in more detail in section 4.3 below.

The above evaluation results for the inflation densities from the euro-area SPF point to some noticeable limitations in particular as regards their information about specific relatively high or low inflation outcomes. In interpreting the result for the fixed event thresholds at shorter horizons (H = 1), the relatively poor predictive performance compared with the equivalent GDP events warrants particular note. On the one hand, due to more significant data revisions for GDP, one might have expected the inflation probabilities to perform better than the equivalent GDP probabilities. However, this result mirrors previous findings which considered the point forecast performance from the euro-area SPF. As discussed in Genre et al. (2013), an important factor which may explain this is the substantial role for commodity price shocks in driving euro-area inflation over the sample period. As shown in Bowles et al. (2010), the predictive performance of the SPF for core inflation has tended to be somewhat better. In addition, it should be noted that the effective forecast horizon for inflation at H = 1 is longer due to the more significant publication lags for the GDP release.[13] It is also of interest to compare our

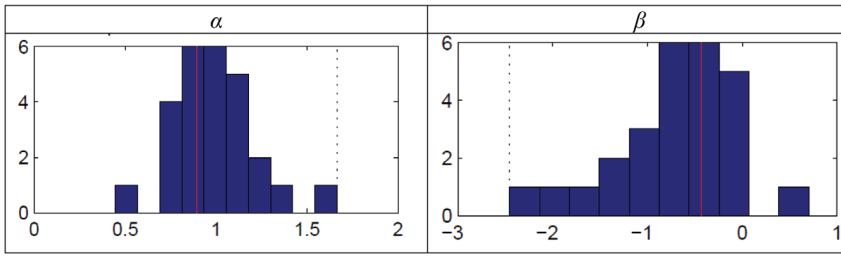
[13] As the ECB SPF sets the horizon exactly one year after the latest publication release, the forecast target date for GDP is two quarters from the survey quarter, while for inflation it is eleven months.

findings for inflation with some of the existing empirical literature that has analyzed the density forecasts from the SPF for the United States. For example, Clements (2006) used the probability integral transform method and finds that the efficiency of the inflation densities from the U.S. SPF is often rejected. However, echoing our more positive results in relation to the direction-of-change forecasts, he does find that the aggregate inflation densities from the U.S. SPF are conditionally efficient compared with a naïve "no change" prediction. Diebold, Tay and Wallis (1999) have also used the probability integral transform (PIT) to evaluate the inflation densities from the U.S. SPF and conclude that the U.S. SPF densities for inflation are not optimal, as the PITs are non-uniform and serially correlated. Part of the reason for poor density performance may relate to overconfidence in the inflation densities as highlighted for the euro area in Kenny, Kostka and Masera (2014a), a feature also documented by Giordani and Söderlind (2006).

4.3 Exploring Further the 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 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 either when aggregated or at a micro level. For most of the events that we have examined, our analysis has supported the pooling of the individual densities across forecasters in terms of both the estimated slope and the intercept coefficients in the panel regressions. In addition, we are often unable to reject the hypothesis of a common forecast error variance among individual forecasters. An exception was the inflation density forecasts, where the evidence pointed to more statistically significant differences in individual forecast performance, as reflected in heterogeneous slope and intercept coefficients. In this section we explore further the extent and nature of heterogeneity in SPF density forecasts at an individual level. To shed light on this, we report the results from the estimation of equation (3) but allow for variation in the intercept and slope parameters across individuals. As an example, figure 5 provides information

Figure 5. Histogram of Individual-Level Parameter Estimates: Inflation > 2% (H = 2)



Notes: 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 panel regressions. The dotted line denotes the estimated parameters based on the probabilities from the aggregate distributions.

on the estimated constant and slope parameter at the individual level for the case of high inflation outcomes above 2 percent at the longer horizon (H = 2). It can be recalled from table 8 that for this event the commonality of parameters across individuals was rejected. The figure depicts a histogram measuring on the vertical axis the number of individual forecasters, with the estimated parameter 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 table 7) 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 underpredict the occurrence of relatively high inflation outcomes. In terms of signaling power (resolution), the individual-level parameter estimates for β also suggest substantial heterogeneity, with the estimates ranging from below –2 to a small positive value in the case of the forecaster with the highest resolution. Yet, despite the heterogeneity in individual forecast attributes, all forecasters appear to be far away from having well-calibrated probability assessments for this event. A more formal assessment of this is provided below, where we report the share of individual

forecasters for which our different hypotheses on individual forecast attributes can be rejected.

Table 9 summarizes 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 rejection rates measured as the number of individuals for which the hypothesis is rejected (at the 10 percent level) expressed as a share of the total number of individuals in the panel.[14] Hence, the table provides some summary information on the degree of heterogeneity in density forecast performance. Rejection rates at the upper and lower bounds—i.e., that are close to either 0 percent or 100 percent—indicate a high level of homogeneity in forecast performance, while shares that are away from these bounds highlight possible heterogeneity in forecaster performance. The figures in table 9 confirm that the previous results concerning the relatively well-calibrated predictions for the fixed threshold event of GDP growth < 1 percent at the short horizon (H = 1) and the direction-of-change forecasts at both horizons are relatively widely shared across individual forecasters. For example, for the GDP growth < 1 percent event at the short horizon (H = 1), we can reject the hypothesis of perfectly calibrated forecasts for only 13 percent of the forecasters in our panel. Similarly, the miscalibration of the fixed threshold event of GDP growth > 2 percent at longer horizons (H = 2) is widely shared across many forecasters in the panel, as indicated by rejection rates of 88 percent and 76 percent for the high and low outcomes, respectively. At the same time, some regressions exhibit a somewhat higher degree of heterogeneity. For instance, the hypothesis of perfect calibration of the GDP growth > 2 percent event forecast at the short horizons (H = 1) was neither rejected in the aggregate nor in the pooled approach (see tables 4 and 5, respectively). However, the same hypothesis can be rejected for 29 percent of all individual forecasters. Regarding inflation, we also observe that the miscalibration of the above 2 percent event is quite homogenous across all forecasters at both

[14]The rejection rates are computed after correcting for the false discovery rate in sequential hypothesis testing as proposed by Benjamini and Hochberg (1995). In this setup, the null is rejected at the $\alpha^{\text{cor}} = 10\% * i/N$ level, i being the individual with the i-th lowest p-value from the set of individual-level regressions.

Table 9. Individual-Level Tests of Risk Forecasts (percentage of individuals for which hypothesis is rejected)

	$\mathbf{H_0}$: $\boldsymbol{\alpha = 0, \beta = 1}$		$\mathbf{H_0}$: $\boldsymbol{\beta \leq 0}$		$\mathbf{H_0}$: $\boldsymbol{\gamma \leq 0}$	
	$\mathbf{H = 1}$	$\mathbf{H = 2}$	$\mathbf{H = 1}$	$\mathbf{H = 2}$	$\mathbf{H = 1}$	$\mathbf{H = 2}$
GDP Growth $> 2\%$	29%	88%	54%	0%	21%	24%
GDP Growth $< 1\%$	13%	76%	96%	0%	42%	20%
Higher GDP Growth	21%	4%	96%	100%	75%	60%
Inflation $> 2\%$	88%	100%	4%	4%	8%	44%
Inflation $< 1\%$	63%	40%	8%	0%	21%	8%
Higher Inflation	33%	72%	92%	92%	88%	80%

Notes: The table reports the share of individuals for which the indicated hypotheses are rejected (at the $\alpha = 10\%$ level) expressed as a share of the total number of individuals in the panel. The estimated rejection rates include the correction for the false discovery rate in sequential hypothesis testing proposed by Benjamini and Hochberg (1995). In this setup, the null is rejected at the $\alpha^{\text{cor}} = 10\% * i/N$ level, i being the individual with the i-th lowest p-value from the set of individual-level regressions.

horizons, as indicated by rejection rates of 88 percent (H = 1) and 100 percent (H = 2). However, for the inflation outcome below 1 percent and the direction-of-change forecasts, it is interesting to observe that the hypothesis of well-calibrated forecasts is only rejected for a smaller but non-negligible share of forecasters. This points to a stronger degree of heterogeneity in forecaster skill for these specific events.

Table 9 also provides information on the resolution and signaling 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 percent level), again expressed as a share of the total number of individuals. Also reported are the equivalent rejection rates using the Pesaran and Timmermann (2009) adjustment for the clustering in the binary outcome variable. Rejection of these hypotheses implies that the probability forecasts are informative in the sense that they have non-zero resolution *and* are positively correlated with the event's occurrence. Concerning this important attribute of the SPF probability forecast, we again observe a high degree of homogeneity across individuals with rejection rates that are close to the upper and lower bounds of 100 percent and 0 percent. The previous findings of positive signaling power for the direction-of-change forecasts for both GDP growth and inflation are shown to be very homogenous and shared by most or even all experts in the panel (for both horizons, the rejection rates are above 90 percent for the hypothesis that $\beta \leq 0$). Equally, the lack of any positive signaling power for the fixed threshold inflation events is shared by nearly all forecasters for both horizons (as indicated by rejection rates that are below 10 percent). This picture is somewhat modified by the results of the Pesaran and Timmermann test for which the rejection rates are further away from the 100 percent and 0 percent bounds. For instance, we find lower rejection rates for the short-horizon GDP forecasts compared with the equivalent test for zero or negative resolution ($\beta \leq 0$). These conflicting indications of potential significant heterogeneity in the SPF forecasts may reflect the small sample that is currently available for empirical analysis. For example, in the case of the Pesaran and Timmermann test, four parameters need to be estimated for each individual, with a maximum of fifty-four observations. Nonetheless, the evidence of more significant heterogeneity in individual forecast performance

will warrant further investigation in the future particularly as more data become available.

The above analysis of density forecast heterogeneity demonstrates that when such forecasts are poorly calibrated or noninformative, this poor performance often tends to be widely shared across the sample of forecasters in our panel. Correspondingly, 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 high and low GDP growth outcomes but only at short horizons—such forecaster skill is often quite evenly distributed across most individuals. At the same time, the extent of heterogeneity in individual forecaster skill can also depend on the specific forecasting task at hand. For some specific events and horizons, and particularly for the calibration of the inflation forecasts, we observe rejection rates that are suggestive of more noteworthy heterogeneity. This evidence on forecaster heterogeneity strikes a note of caution for the strong poolability that was observed in many of the panel regressions. It also accords more closely with other studies of the U.S. SPF such as Engleberg, Manski, and Williams (2009) and previously Davies and Lahiri (1995) which, although using different methods and perspectives, had emphasized the heterogeneity in individual SPF replies. Future research might therefore consider carefully 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, weak signaling power, and low information content.[15]

[15]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 (2012) have proposed methods to estimate optimal combination weights and have applied their method to the euro-area SPF, finding some modest gains.

5. Conclusions

In this paper we have attempted to shed further light on the information content of the density forecasts of macroeconomists collected in surveys. 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, and bearing a small-sample caveat in mind, we have observed relatively low information content in the SPF density forecasts for relatively high and low outcome events. 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 confirms the ability of the 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. In general, our analysis also points to a rather high degree of homogeneity in forecaster skill along some of the dimensions that we have examined and when assessed from an ex post perspective. At the same time, we have uncovered some evidence to suggest that the heterogeneity in forecaster skill can also depend on the specific forecasting task. Future research might therefore consider carefully the extent to which such heterogeneity can be exploited in order to improve the usefulness of surveys such as the SPF.

Appendix 1

In line with the discussion in section 2.2, we construct our variance-covariance matrix under the following simplifying assumptions:

$$E[\varepsilon_{i,t+l}\varepsilon_{j,t+m}] = \left\{ \begin{array}{ll} \sigma_{i|l-m|} \geq 0 & for\ i = j\ and\ |l-m| \leq \tau \\ \delta_{l-m|} \geq 0 & for\ i \neq j\ and\ |l-m| \leq \tau \\ 0 & otherwise \end{array} \right\}. \tag{6}$$

In (6), both individual-specific error variances ($\sigma_{i,0}$) and the covariances of the lagged errors up to order $\tau(\sigma_{i,1}, \ldots, \sigma_{i,\tau})$ are allowed to be different across each individual in the SPF panel. In contrast,

$\delta_0, \ldots, \delta_\tau$ reflect coincident and lagged covariances across individuals that are due to aggregate shocks and, hence, are common to all individual forecasters. The corrected estimate of $\hat{\Omega}$ thus has several off-diagonal non-zero elements capturing the heterogeneity as well as the time and cross-sectional correlation in the residuals of equation (2). More specifically, it takes the form given in equation (7).

$$\hat{\Omega} = \begin{bmatrix} A_1 & B & \cdot & \cdot & \cdot & B \\ B & A_2 & B & \cdot & & \\ B & B & A_3 & B & \cdot & \\ \cdot & \cdot & \cdot & & & \\ B & \cdot & \cdot & \cdot & & A_N \end{bmatrix} \tag{7}$$

For instance, in the case of the four-quarter-ahead probability forecast ($\tau = 4$), the matrices A and B are constructed as follows:

$$A_i = \begin{bmatrix} \sigma_{i,0} & \sigma_{i,1} & \sigma_{i,2} & \sigma_{i,3} & \sigma_{i,4} & 0 & \cdots & 0 \\ \sigma_{i,1} & \sigma_{i,0} & \sigma_{i,1} & \sigma_{i,2} & \sigma_{i,3} & \sigma_{i,4} & \cdots & 0 \\ \sigma_{i,2} & \sigma_{i,1} & \sigma_{i,0} & \sigma_{i,1} & \sigma_{i,2} & \sigma_{i,3} & \cdots & 0 \\ \sigma_{i,3} & \sigma_{i,2} & \sigma_{i,1} & \sigma_{i,0} & \sigma_{i,1} & \sigma_{i,2} & \cdots & 0 \\ \sigma_{i,4} & \sigma_{i,3} & \sigma_{i,2} & \sigma_{i,1} & \sigma_{i,0} & \sigma_{i,1} & \cdots & 0 \\ 0 & \sigma_{i,4} & \sigma_{i,3} & \sigma_{i,2} & \sigma_{i,1} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \sigma_{i,1} \\ 0 & 0 & 0 & 0 & 0 & \cdots & \sigma_{i,1} & \sigma_{i,0} \end{bmatrix} \tag{8}$$

$$B = \begin{bmatrix} \delta_0 & \delta_1 & \delta_2 & \delta_3 & \delta_4 & 0 & \cdots & 0 \\ \delta_1 & \delta_0 & \delta_1 & \delta_2 & \delta_3 & \delta_4 & \cdots & 0 \\ \delta_2 & \delta_1 & \delta_0 & \delta_1 & \delta_2 & \delta_3 & \cdots & 0 \\ \delta_3 & \delta_2 & \delta_1 & \delta_0 & \delta_1 & \delta_2 & \cdots & 0 \\ \delta_4 & \delta_3 & \delta_2 & \delta_1 & \delta_0 & \delta_1 & \cdots & 0 \\ 0 & \delta_4 & \delta_3 & \delta_2 & \delta_1 & \ddots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \delta_1 \\ 0 & 0 & 0 & 0 & 0 & \cdots & \delta_1 & \delta_0 \end{bmatrix}. \tag{9}$$

Using (7), it is then possible to derive the GMM estimates for α and β as described in section 2.2 and to draw inference on the values of these parameters that is robust even in the presence of aggregate

shocks and individual-specific error variances. The elements of (8) and (9) are then computed as in (10)–(12).

$$\sigma_{i,0} = T^{-1} \sum_{t=1}^{T} \varepsilon_{i,t}^2 \tag{10}$$

$$\sigma_{i,|l-m|} = [T - |l-m|]^{-1} \sum_{t=1}^{T-|l-m|} \varepsilon_{i,t}\varepsilon_{i,t+|l-m|}$$

$$for \quad |l-m| = 0, ., ., \tau \tag{11}$$

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

$$for \quad |l-m| = 0, ., ., \tau \tag{12}$$

Appendix 2. Sensitivity to Choice of Event Thresholds

Our analysis of the individual and aggregate SPF probability forecasts for fixed threshold events has been based on upper and lower thresholds of 1 percent and 2 percent. The choice of these thresholds was primarily motivated by their interest from a user perspective. For example, inflation above 2 percent is seen as inconsistent with the ECB definition of price stability, while rates that are below 1 percent are often seen as raising concerns about disinflationary or even deflationary pressure. Notwithstanding the interest in these thresholds from an economic perspective, it is interesting to assess the sensitivity of our broad conclusions to the precise numerical choice of threshold values. Such a sensitivity analysis is conducted for the aggregate- and individual-level analysis in tables 10 and 11, respectively. In particular, these tables report the p-values corresponding to the tests of well-calibrated forecasts and positive resolution whilst varying the upper threshold over the interval [1.75, 2.00, 2.25] and the lower threshold over the interval [0.75, 1.0, 1.25]. Overall, the sensitivity analysis reported in tables 10 and 11 highlights that our main results appear quite stable when we vary the numerical thresholds for the fixed threshold events. At the 10 percent level, rarely do we observe a change in the result of the hypotheses test when the

Table 10. Sensitivity of Inference to Variation of Event Thresholds: Aggregate Forecast (event threshold $+/-0.25\%$)

A. Event 1 (Target Variable $> x\%$)						
	$x = 1.75$		$x = 2.0$		$x = 2.25$	
	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$
GDP Growth $> x\%$ (H = 1)	0.51	0.00	0.33	0.00	0.42	0.00
GDP Growth $> x\%$ (H = 2)	0.00	0.74	0.00	0.79	0.00	0.60
Inflation $> x\%$ (H = 1)	0.00	0.13	0.00	0.61	0.01	0.36
Inflation $> x\%$ (H = 2)	0.00	0.99	0.00	1.00	0.00	0.99
B. Event 2 (Target Variable $< x\%$)						
	$x = 0.75$		$x = 1.0$		$x = 1.25$	
	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$
GDP Growth $< x\%$ (H = 1)	0.22	0.00	0.24	0.00	0.44	0.00
GDP Growth $< x\%$ (H = 2)	0.07	0.62	0.04	0.66	0.01	0.81
Inflation $< x\%$ (H = 1)	0.02	0.30	0.00	0.60	0.06	0.32
Inflation $< x\%$ (H = 2)	0.00	0.92	0.00	0.93	0.00	0.95

Table 11. Sensitivity of Inference to Variation of Event Thresholds: Pooled Sample (event threshold $+/-0.25\%$)

A. Event 1 (Target Variable $> x\%$)						
	$x = 1.75$		$x = 2.0$		$x = 2.25$	
	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$
GDP Growth $> x\%$ (H = 1)	0.10	0.01	0.05	0.01	0.12	0.02
GDP Growth $> x\%$ (H = 2)	0.00	0.74	0.00	0.80	0.00	0.54
Inflation $> x\%$ (H = 1)	0.00	0.02	0.00	0.41	0.00	0.24
Inflation $> x\%$ (H = 2)	0.00	0.99	0.00	1.00	0.00	0.99
B. Event 2 (Target Variable $< x\%$)						
	$x = 0.75$		$x = 1.0$		$x = 1.25$	
	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$	H_0: $\alpha = 0, \beta = 1$	H_0: $\beta \leq 0$
GDP Growth $< x\%$ (H = 1)	0.14	0.00	0.14	0.00	0.20	0.00
GDP Growth $< x\%$ (H = 2)	0.00	0.63	0.00	0.69	0.00	0.78
Inflation $< x\%$ (H = 1)	0.01	0.21	0.00	0.35	0.00	0.17
Inflation $< x\%$ (H = 2)	0.00	0.88	0.00	0.91	0.00	0.94

thresholds are varied in this way. One exception is that the hypothesis of a negative β tends to be rejected when the threshold for the high inflation outcome (H = 1) is lowered from 2 percent to 1.75 percent in the panel regressions (table 11). Lowering the threshold for this event thus casts the inflation probability forecasts in a somewhat better light. At the same time, however, it can be observed from table 11, panel A, that the hypothesis that these probability forecasts are perfectly calibrated is rejected for all three fixed thresholds.

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 (1): 289–300.

Berkowitz, J. 2001. "Testing Density Forecasts with Applications to Risk Management." *Journal of Business and Economic Statistics* 19 (4): 465–74.

Boero, G., J. Smith, and K. F. Wallis. 2008. "Evaluating a Three-Dimensional Panel of Point Forecasts: The Bank of England Survey of External Forecasters." *International Journal of Forecasting* 24 (3): 354–67.

———. 2011. "Scoring Rules and Survey Density Forecasts." *International Journal of Forecasting* 27 (2): 379–93.

Bowles, C., R. Friz, V. Genre, G. Kenny, A. Meyler, and T. Rautanen. 2010. "An Evaluation of the Growth and Unemployment Forecasts in the ECB SPF." *Journal of Business Cycle Measurement and Analysis* 2010 (2): 63–90.

Brier, G. W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Christoffersen, P. F. 1998. "Evaluating Interval Forecasts." *International Economic Review* 39 (4): 841–62.

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 (1): 49–64.

———. 2010. "Explanations of Inconsistencies in Survey Respondents' Forecasts." *European Economic Review* 54 (4): 536–49.

Clements, M. P., F. Joutz, and H. O. Stekler. 2007. "An Evaluation of the Forecasts of the Federal Reserve: A Pooled Approach." *Journal of Applied Econometrics* 22 (1): 121–36.

Conflitti, C., C. De Mol, and D. Giannone. 2012. "Optimal Combination of Survey Forecasts." ECARES Working Paper No. 2012-023, Universite Libre de Bruxelles.

Croushore, D. 2010. "An Evaluation of Inflation Forecasts from Surveys Using Real-Time Data." *BE Journal of Macroeconomics: Contributions* 10 (1, Article 10): 1–32.

D'Agostino, A., K. McQuinn, and K. Whelan. 2012. "Are Some Forecasters Really Better Than Others?" *Journal of Money, Credit and Banking* 44 (4): 715–32.

Davies, A., and K. Lahiri. 1995. "A New Framework for Analyzing Survey Forecasts Using Three-Dimensional Panel Data." *Journal of Econometrics* 68 (1): 205–27.

Dawid, A. P. 1984. "Statistical Theory: The Prequential Approach." *Journal of the Royal Statistical Society: Series A* 147 (2): 278–90.

Diebold, F. X., and G. D. Rudebusch. 1989. "Scoring the Leading Indicators." *Journal of Business* 62 (3): 369–91.

Diebold, F. X., A. S. Tay, and K. F. Wallis. 1999. "Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters." In *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W. J. Granger*, ed. R. Engle and H. White, 76–90 (chapter 3). Oxford: Oxford University Press.

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., and S. van Norden. 2011. "Kernel-Based Calibration Diagnostics for Inflation and Recession Probability Forecasts." *International Journal of Forecasting* 27 (4): 1041–57.

———. 2012. "Assessing Gross Domestic Product and Inflation Probability Forecasts Derived from Bank of England Fan Charts." *Journal of the Royal Statistical Society: Series A* 175 (3): 713–27.

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

Genre, V., G. Kenny, A. Meyler, and A. Timmermann. 2013. "Combining the Expert Forecasts: Can Anything Beat the Simple Average?" *International Journal of Forecasting* 29 (1): 108–21.

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): 1027–43.

Gneiting, T., F. Balabaoui, and A. E. Raftery. 2007. "Probabilistic Forecasts, Calibration and Sharpness." *Journal of the Royal Statistical Society: Series B* 69 (2): 243–68.

Granger, C. W. J., and M. H. Pesaran. 2000. "Economic and Statistical Measures of Forecast Accuracy." *Journal of Forecasting* 19 (7): 537–60.

Harding, D., and A. Pagan. 2010. "Can We Predict Recessions?" NCER Working Paper No. 69 (December).

Holden, K., and D. A. Peel. 1990. "On Testing for Unbiasedness and Efficiency of Forecasts." *The Manchester School* 58 (2): 120–27.

Hsiao, C. 2003. *Analysis of Panel Data*. Econometric Society Monographs, Book 34. Cambridge University Press.

Jore, A. S., J. Mitchell, and S. P. Vahey. 2010. "Combining Forecast Densities from VARs with Uncertain Instabilities." *Journal of Applied Econometrics* 25 (4): 621–34.

Keane, M. P., and D. E. Runkle. 1990. "Testing the Rationality of Price Forecasts: New Evidence from Panel Data." *American Economic Review* 80 (4): 714–35.

Kenny, G., T. Kostka, and F. Masera. 2014a. "Density Characteristics and Density Forecast Performance: A Panel Analysis." *Empirical Economics* 48 (3): 1203–31.

———. 2014b. "How Informative Are the Subjective Density Forecasts of Macroeconomists?" *Journal of Forecasting* 33 (3): 163–85.

Lahiri, K., and J. G. Wang. 2006. "Subjective Probability Forecasts for Recessions: Evaluation and Guidelines for Use." *Business Economics* 41 (2): 26–37.

———. 2007. "The Value of Probability Forecasts as Predictors of Economic Downturns." *Applied Economics Letters* 14 (1): 11–14.

———. 2013. "Evaluating Probability Forecasts for GDP Declines Using Alternative Methodologies." *International Journal of Forecasting* 29 (1) 175–90.

Mincer, J., and V. Zarnowitz. 1969. "The Evaluation of Economic Forecasts." In *Economic Forecasts and Expectations*, ed. J. Mincer. New York: National Bureau of Economic Research.

Mitchell, J., and K. Wallis. 2011. "Evaluating Density Forecasts: Forecast Combinations, Model Mixtures, Calibration and Sharpness." *Journal of Applied Econometrics* 26 (6): 1023–40.

Murphy, A. H. 1973. "A New Vector Partition of the Probability Score." *Journal of Applied Meteorology* 12 (4): 595–600.

———. 1988. "Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient." *Monthly Weather Review* 116 (12): 2417–24.

Murphy, A. H., and E. S. Epstein. 1967. "A Note on Probability Forecasts and 'Hedging'." *Journal of Applied Meteorology* 6 (6): 1002–04.

Murphy, A. H., and R. L. Winkler. 1992. "Diagnostic Verification of Probability Forecasts." *International Journal of Forecasting* 7 (4): 435–55.

Newey, W. K., and K. D. West. 1987. "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55 (3): 703–8.

Pesaran, H. M., and A. Timmermann. 2009. "Testing Dependence among Serially Correlated Multicategory Variables." *Journal of the American Statistical Association* 104 (485): 325–37.

Tay, A. S., and K. F. Wallis. 2000. "Density Forecasting: A Survey." *Journal of Forecasting* 19 (4): 235–54. Reprinted in *A Companion to Economic Forecasting,* ed. M. P. Clements and D. F. Hendry, 45–68. Oxford: Blackwell, 2002.

Yates, J. F. 1982. "External Correspondence: Decompositions of the Mean Probability Score." *Organizational Behavior and Human Performance* 30 (1): 132–56.