University of Reading

Are macroeconomic density forecasts informative?

Article

Accepted Version

Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0

Clements, M. ORCID: https://orcid.org/0000-0001-6329-1341
(2018) Are macroeconomic density forecasts informative? International Journal of Forecasting, 34 (2). pp. 181-198. ISSN 0169-2070 doi: https://doi.org/10.1016/j.ijforecast.2017.10.004 Available at https://centaur.reading.ac.uk/72924/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1016/j.ijforecast.2017.10.004

Publisher: Elsevier

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Are Macroeconomic Density Forecasts Informative?

Michael P. Clements*

ICMA Centre,

Henley Business School,

University of Reading,

Reading RG6 6BA

October 3, 2017

Abstract

We consider whether survey density forecasts (such as the inflation and output growth histograms of the US Survey of Professional Forecasters) are superior to unconditional density forecasts. The unconditional forecasts assume that the average level of uncertainty experienced in the past will prevail in the future, whereas the SPF projections ought to be adapted to current conditions and the outlook at each forecast origin. The SPF forecasts might be expected to outperform the unconditional densities at the shortest horizons, but this does not transpire to be the case for the aggregate forecasts of either variable, or for the majority of the individual respondents for forecasting inflation.

Keywords: probability distribution forecasts, aggregation, Kullback-Leibler information criterion.

JEL classification: C53.

*I am grateful to two anonymous referees, an Associate Editor, and Editor Dick van Dijk for helpful comments, as well as to conference participants at the International Symposium on Forecasting, Santander, 2106

1 Introduction

There has been much interest in survey forecasts in recent years, driven in part by the opportunities they offer to test theories of expectations-generation (see, e.g., Pesaran and Weale (2006), Coibion and Gorodnichenko (2012, 2015) and Andrade and Le Bihan (2013), amongst many others) and for improved forecasting accuracy, either as direct forecasts themselves (see, e.g., Ang, Bekaert and Wei (2007), Clements (2015)) or as an adjunct to other forecasting models (see, e.g., Wright (2013)). As well as collecting point predictions, some surveys elicit respondents' subjective probability distributions, in the form of histograms, offering the promise of 'direct' measures of forecast uncertainty (see, e.g., Giordani and Söderlind (2003)), Rich and Tracy (2010) and Clements (2014)) as an alternative to less theoretically satisfactory measures such as forecaster disagreement, as given by some measure of the cross-sectional dispersion of the point predictions (Zarnowitz and Lambros (1987)).

Our interest is in whether survey respondents are able to form probability assessments about the future values of key macro-variables (such as output growth and inflation) which are more accurate than 'unconditional' benchmark densities. Little is known about how the information content of the subjective probability assessments varies with the forecast horizon: one might surmise that survey forecasters would outperform the benchmarks at short horizons, but any advantage would dissipate as the horizon increases, but (to the best of my knowledge) there is little evidence as to whether this is this case. Of interest is the performance of the aggregate distributions (i.e., averaging across individual respondents), as well as the individual forecasters' assessments, i.e., whether combination (or aggregation) plays an important role.

We consider the US Survey of Professional Forecasters (SPF). We regard the SPF densities as adding value if they are more accurate than the benchmarks, at least at short horizons. We first consider *truly* unconditional density forecasts as the benchmarks, whereby we assume normality and estimate the mean and variance from the historical forecasts. However, these densities are resoundingly rejected both when we test whether they are correctly specified, and when we compare them against the SPF densities, simply because the unconditional mean is a poor estimate of the conditional mean. Hence the rejection of the truly unconditional densities is not surprising, and they constitute an insufficiently challenging benchmark for the SPF densities. We then refine the benchmark forecasts to provide a stiffer challenge: the forecast densities are centred on the median point predictions (of the SPF respondents), and hence draw on forecast origin information, but the scale or dispersion is calculated from the historical variance of the forecast errors, as before. Comparison of the SPF densities to these benchmarks serve to ask whether the SPF densities contain any useful information about the uncertainty or likely dispersion of future outcomes. That is, we shift the focus to second moments, having acknowledged that survey forecasters are able to forecast first moments (as found by, e.g., Ang *et al.* (2007)). [1]

[1] Knüppel and Schultefrankenfeld (2012) are primarily interested in assessing the informativeness of predictions of third moments (.i.e, skewness) by Central Banks. Our focus is on whether the second moment assessments are reasonably accurate as a precursor to the consideration of higher moments.

One might expect the SPF forecasts to outperform the unconditional densities at the shortest horizons assuming the variances of the densities change over time in a way which is at least in part predictable. However, the relative improvements would be expected to diminish as the forecast horizon lengthens, as the role of current conditions in predicting future developments lessens. Our results suggest the opposite: the aggregate and individual histograms are rejected in favour of the benchmark densities at the shorter horizons, reflecting the under-confidence of the survey respondents at within-year horizons, as documented by Clements (2014). That is, the survey respondents tend to over-estimate the degree of uncertainty they face when forecasting at the shorter horizons. We show that this is true at the aggregate level and also holds for individuals. Moreover, at least at the level of the aggregate histograms the mis-specification is found to be systematic, and 'future' densities can be improved using a simple correction calculated from an in-sample or training set.

We should emphasize that the finding of under-confidence runs counter to the prevailing view in the literature on behavioural economics and finance (see, e.g., the surveys by Rabin (1998) and Hirshleifer (2001)). However, it is a much older view. For example, in discussing over-confidence, Malmendier and Taylor (2015) refer to Smith (1776, Book 1, Chapter X), 'The over-weening conceit which the greater part of men have in their own abilities, is an ancient evil remarked by the philosophers and moralists of all ages'. Hence the findings we present are in some ways a challenge, and are presented in the hope that they may foster further work in this area.

Our contribution does not stand alone - there has been earlier work. Giordani and Söderlind (2003) find that SPF respondents' confidence intervals for annual inflation one-year ahead have actual coverage rates markedly lower than the nominal, indicating over-confidence, but these authors do not consider shorter horizons. Giordani and Söderlind (2006) consider the US SPF real annual GDP (and GNP) forecasts 1982-2003, and consider forecasts made in each of the four quarters of the year of the current year annual growth rate (i.e., forecasts from 1-quarter to 1-year ahead, approximately). However, they report coverage rates for all the four horizons taken together. Kenny, Kostka and Masera (2012) consider the ECB's SPF and find over-confidence in the respondents' Euro area GDP growth and inflation forecasts, but these are of one and two-years ahead forecast horizons. Clements (2014) compares the *ex ante* uncertainty estimates of the SPF respondents (that is, uncertainty estimates calculated from their histograms) to *ex post* estimates, and finds under-confidence, and Clements and Galvão (2017) compare the survey estimates to model-based estimates. Both find under-confidence on the part of the shorter-horizon survey estimates. Clements (2014) essentially compares the actual forecast errors with those expected based on the *ex ante* assessments. The comparisons reported in Clements and Galvão (2017) show that econometric models give more accurate estimates of uncertainty than the survey forecasts. Their study is real-time, in the sense that the models' datasets match the data available to the survey respondents at each forecast origin, and so do not benefit from a 'look-forward' bias. The models are specified and estimated only on data the survey respondents would have had access to. Nevertheless, the failure of the survey respondents not to have used the modern econometric

modelling techniques of Clements and Galvão (2017) is not surprising. Whereas being out-performed by the benchmarks we use in these paper might call into question the value of the density forecasts.

We contribute to the literature on survey expectations in a number of ways: by comparing the SPF histograms directly to unconditional, empirical distributions, we assess the value of the forecast horizon 'conditioning' information; we use different ways of assessing and comparing the SPF and benchmark forecasts; we consider whether survey forecasters are more skilful at assessing the probabilities attached to particular regions of density (corresponding to events of particular interest); and we consider whether simple mechanistic corrections of the survey forecasts improves their accuracy.

Our empirical investigation considers both whether the SPF densities and the benchmark densities are correctly specified, and provides a comparison of the two, not requiring that either set closely approximates the truth. We are careful to check that our findings are not dependent on changes in the way the survey has been implemented over time, or on any mismatch between point predictions and histogram means (e.g., Engelberg, Manski and Williams (2009)), and we consider alternative loss functions.

Figures 1 and 2 present a selective look-ahead to our results. For annual output growth and inflation, respectively, they present time series of the aggregate densities and the outcomes (advance estimates) for i) top panels, the year-ahead forecasts, made in response to the surveys held in the first quarters of the years 1982 to 2013, and ii) bottom panels, the one quarter-ahead forecasts, made in response to the surveys in the fourth quarters 1981 to 2013. Simply eye-balling the densities and the associated outcomes suggests short horizon forecasts are too dispersed: realizations outside the 80% interval (defined by the 90th and 10th percentiles) should occur a fifth of the time, but there are no such instances for output growth or inflation. Moreover, realizations outside the interquartile range should occur half the time. For output growth, from around 2000 onwards it appears that the actuals are well within the interquartile range much of the time, and for inflation occurrences outside the IQ range are rare. There are limitations to what can be learnt from a graphical analysis, and in the paper we provide formal assessments, and comparisons to benchmark forecasts, as well as an analysis of the individual forecasters.

We regard the value of survey macro-forecasts as established in the case of first-moment prediction, but as unproven in terms of the probability assessments implied by the histogram forecasts.

The remainder of the paper is as follows. Section 2 reviews the literature on forecast density evaluation, and on comparisons between forecast densities. Section 3 describes the survey data. Section 4 describes the construction of the benchmark densities, used to gauge the value of conditioning on forecast origin information. The results are given in section 5. Section 6 assesses the robustness of our main findings to the assumptions which have been made, and includes a section considering a number of alternative density scoring rules. Section 7 considers a simple correction based on the past performance of the SPF densities, which delivers more accurate densities over the out-of-sample period. Section 8 concludes.

2 The Evaluation of Survey Density Forecasts, and Comparisons to Benchmarks

A popular way of evaluating survey density forecasts is based on the probability integral transform, dating back at least to Rosenblatt (1952), with recent contributions by Shephard (1994), Kim, Shephard and Chib (1998) and Diebold, Gunther and Tay (1998). Diebold *et al.* (1998) and Granger and Pesaran (2000) show that a density forecast that coincides with the 'true' forecast density will be optimal in terms of minimizing expected loss irrespective of the form of the (generally unknown) user's loss function. In some applications only a portion of the forecast density may be relevant, as for example in financial risk management, where the tail quartile of the expected distribution of returns plays a prominent role (in Value at Risk calculations), or in macro inflation forecasting, where the focus is on the probability that inflation will exceed a target value. Tools have been developed for the study of quartiles and for events derived from density forecasts (see, e.g., Engle and Manganelli (2004), Clements (2004), Gneiting and Ranjan (2011) and Diks, Panchenko and van Dijk (2011)). Our focus is on the whole density but we also look at a particular region of the density, which correspond to lower than normal growth of real GDP, and rates of inflation close to the target value of 2%.

There are two parts to our empirical investigation of the forecast densities. In the first part, we assess how well the SPF forecast densities approximate the true, unknown densities, using the probability integral transform, and in particular, the extension due to Berkowitz (2001). In the second, we compare the SPF densities to the benchmark densities. Hence the SPF densities may be mis-specified - of interest is whether they are nonetheless superior to the benchmarks. These comparisons are motivated by Lee, Bao and Saltoglu (2007) and Mitchell and Hall (2005), and the recognition that the Kullback-Leibler Information Criterion (Kullback and Leibler (1951), KLIC) measure of the divergence between a forecast density and the true density can be adapted to compare two or more densities, without making the assumption that any of the densities is correctly-specified. The KLIC will be used to compare the SPF densities against the benchmark densities in the form of a Diebold and Mariano (1995) test of equal predictive ability, and amounts to a comparison in terms of logarithmic score. The testing of rival density forecasts is formalized by Amisano and Giacomini (2007) for a class of scoring rules, and their approach is shown to be valid when the densities are generated from mis-specified models with estimated parameters, and when the resulting densities are dynamically mis-specified. As well as considering the logarithmic score (henceforth log score) which is perhaps the most popular loss function of scoring rule for densities, [2] we assess the robustness of our findings to the quadratic and ranked probability scores.

The benchmark densities are such that the rejection of the null of equal density accuracy in favour of the SPF densities would imply that the conditioning on the forecast origin information implicit in the survey

[2] Winkler (1967) is the classic reference on scoring rules, and more recently Gneiting and Raftery (2007).

forecasts results in the superior accuracy.[3] One would expect the forecast origin information would become less valuable for determining the scale of the forecast density as the forecast horizon increases. By considering density forecasts with horizons from approximately one quarter to one year ahead, it might be possible to determine the horizon at which the survey forecasts become 'uninformative', in the sense that, evaluated by KLIC, the conditioning information yields no improvement in accuracy.[4]

2.1 Density evaluation

Suppose we have a series of 1-step forecast densities for the value of a random variable $\{Y_t\}$, denoted by $p_{Y,t-1}(y)$, where $t = 1, \ldots, n$. The probability integral transforms (p.i.t.'s) of the realizations of the variable with respect to the forecast densities are given by:

$$z_t = \int_{-\infty}^{y_t} p_{Y,t-1}(u)du \equiv P_{Y,t-1}(y_t) \tag{1}$$

for $t = 1, \ldots, n$, where $P_{Y,t-1}(y_t)$ is the forecast probability of Y_t not exceeding the realized value y_t. In terms of the random variables $\{Y_t\}$, rather than their realized values $\{y_t\}$, we obtain random variables denoted by $\{Z_t\}$:

$$Z_t = \int_{-\infty}^{Y_t} p_{Y,t-1}(u)du \equiv P_{Y,t-1}(Y_t).$$

When the forecast density equals the true density, $f_{Y,t-1}(y)$, it follows by a simple change-of-variables argument that $Z_t \sim U(0,1)$, where $U(0,1)$ is the uniform distribution over $(0,1)$. Even though the actual conditional densities may be changing over time, provided the forecast densities match the actual densities at each t, then $Z_t \sim U(0,1)$ for each t, and the Z_t are independently distributed of each other, such that the realized time series $\{z_t\}_{t=1}^n$ is an iid sample from a $U(0,1)$ distribution.

This suggests we can evaluate whether the conditional forecast densities match the true conditional densities by testing whether $\{z_t\}_{t=1}^n$ is iid$U(0,1)$. Berkowitz (2001) suggested taking the inverse normal CDF transformation of the $\{z_t\}_{t=1}^n$ series, to give, say, $\{z_t^*\}_{t=1}^n$, on the grounds that more powerful tools can be applied to testing the null that the $\{z_t^*\}_{t=1}^n$ are iid$N(0,1)$ (for $h = 1$) compared to one of iid uniformity of the original $\{z_t\}_{t=1}^n$ series. He proposes a one-degree of freedom test of independence against a first-order autoregressive structure, as well as a three-degree of freedom test of zero-mean, unit variance and independence. In each case the maintained assumption is that of normality, so that standard likelihood ratio tests are constructed using the gaussian likelihoods.

[3] In a similar vein, for point forecasts Diebold and Kilian (2001) suggest measuring predictability by comparing the expected loss of a short-horizon forecast to a long-horizon forecast. We use the unconditional density as the 'long-horizon' forecast, against which the SPF density forecasts are compared as the 'short-horizon' forecasts. Note that our primary benchmark forecasts are unconditional in terms of the dispersion about the mean, but the mean is conditioned on forecast origin information. As noted in the text, *truly* unconditional benchmarks are found to be clearly inferior.

[4] The way in which this discussion is framed assumes that the one-quarter ahead survey forecasts will outperform, and that 1-year forecasts will be no better, but this does not turn out to be the case.

The Berkowitz (2001) three degree of freedom test is given by:

$$LR_B = -2(L(0,1,0) - L\left(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}\right)) \tag{2}$$

where $L(0,1,0) = \sum_{t=1}^{n}\left[\ln\phi\left(z^*_{t|t-1}\right)\right]$ is the value of the Gaussian log-likelihood for an independently distributed standard normal random variable ($\phi(.)$ is the $N(0,1)$ pdf), and:

$$L\left(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}\right) = \sum_{t=1}^{n}\left[\ln\left(\phi\left(\left(z^*_{t|t-1} - \hat{\mu} - \hat{\rho}z^*_{t-1|t-2}\right)/\hat{\sigma}\right)/\hat{\sigma}\right)\right]$$

is the maximized log-likelihood for an AR(1) with Gaussian errors ($\hat{}$'s on parameters denote maximum likelihood estimates). As noted by Lee *et al.* (2007), the assumptions of a first-order process, and that it is Gaussian, can be generalized: they allow instead a higher-order autoregression, with iid disturbances that follow a semi-non-parametric density function. However, for quarterly macro data relatively short-sample sizes perhaps warrant the simpler assumptions.

Finally, Knüppel (2015) suggests a simple way of testing whether the z^* are standard normal using the non-standardized, non-central (i.e., 'raw') moments of the z^*, and compares this approach with that of Bai and Ng (2005). We report tests of raw moments for the two moments, and for the first four moments.

2.2 Density comparison

Lee *et al.* (2007) show that the Berkowitz test can be interpreted as a particular form of KLIC-based evaluation of a forecast density compared to the true density, and that the KLIC can also be used to compare (two or more) mis-specified identities.

Firstly, comparing a forecast density to the true density. The KLIC is defined as:

$$KLIC_{t|t-h} = E\left[\ln\left(f_{Y,t-h}(y_t)\right) - \ln\left(p_{Y,t-h}(y_t)\right)\right]$$

where the expectation is with respect to the true density, so that:

$$KLIC_{t|t-h} = \int f_{Y,t-h}(y_t)\ln\left\{\frac{\left(f_{Y,t-h}(y_t)\right)}{\left(p_{Y,t-h}(y_t)\right)}\right\}dy_t.$$

Berkowitz (2001, Proposition 2, p.467) shows that $\ln\left(f_{Y,t-h}(y_t)\right) - \ln\left(p_{Y,t-h}(y_t)\right) = \ln q_{Z^*,t-h}(z^*_t) - \ln\phi(z^*_t)$, where $q_{Z^*,t-h}$ is the true (unknown) density of z^*_t and ϕ is the standard normal. The KLIC is estimated as the sample average of $d_{t|t-h} \equiv \ln q_{Z^*,t-h}(z^*_t) - \ln\phi(z^*_t)$ (over t, for a given h), and if we allow that z^*_t is a Gaussian AR(1), we obtain:

$$\begin{aligned}\overline{KLIC}_h &= \frac{1}{n}\sum_{t=1}^{n} d_{t|t-h} = \frac{1}{n}\sum_{t=1}^{n}\left[\ln\left(\phi\left(\left(z^*_{t|t-h} - \hat{\mu} - \hat{\rho} z^*_{t-1|t-h-1}\right)/\hat{\sigma}\right)/\hat{\sigma}\right)\right] - \frac{1}{n}\sum_{t=1}^{n}\left[\ln\phi\left(z^*_{t|t-h}\right)\right] \\ &= (2n)^{-1} LR_B, \end{aligned} \tag{3}$$

where LR_B is given in (2). Hence the KLIC and Berkowitz test are directly related. The assumption that the $z^*_{t|t-h}$ are independent for optimal density forecasts is valid in our framework, as explained below.

The KLIC can also be used as the loss function in a *comparison* of two density forecasts, using the approach to testing equal predictive accuracy of Diebold and Mariano (1995). Letting the loss differential between the KLICs of the two densities be $d_{t|t-h}$, then

$$\begin{aligned} d_{t|t-h} &= \left(\ln\left(f_{Y,t-h}\left(y_t\right)\right) - \ln\left(p^2_{Y,t-h}\left(y_t\right)\right)\right) - \left(\ln\left(f_{Y,t-h}\left(y_t\right)\right) - \ln\left(p^1_{Y,t-h}\left(y_t\right)\right)\right) \\ &= \ln\left(p^1_{Y,t-h}\left(y_t\right)\right) - \ln\left(p^2_{Y,t-h}\left(y_t\right)\right), \end{aligned} \tag{4}$$

where $p^1()$ and $p^2()$ denote rivals sets of forecast densities.

The average loss differential is:

$$\overline{d}_h = \frac{1}{n}\sum_{t=1}^{n} d_{t|t-h}$$

with limiting distribution:

$$\sqrt{n}\left(\overline{d}_h - E\left(d_{t|t-h}\right)\right) \xrightarrow{d} N\left(0, \xi^2\right) \tag{5}$$

where ξ^2 is the limiting variance.

We will use (5) to compare the SPF forecasts (say, $p^1_{Y,t-h}$) against the benchmark forecasts ($p^2_{Y,t-h}$). In which case, under the null of equal accuracy, $E\left(d_{t|t-h}\right) = 0$, and $\xi^2 = \gamma_0 + 2\sum_{j=1}^{\infty}\gamma_j$, with $\gamma_j = E\left(d_{t|t-h} d_{t-j|t-j-h}\right)$, and (5) is simply:

$$\frac{\overline{d}_h}{\xi/\sqrt{n}} \xrightarrow{d} N\left(0,1\right). \tag{6}$$

As noted in section 2, this corresponds to the approach to comparing density forecasts of Amisano and Giacomini (2007). The Amisano and Giacomini (2007) approach allows for comparisons between mis-specified densities. This is important given that both the SPF and benchmark densities are found to be mis-specified at the shorter horizons, and for this reason it might be important to allow for dynamic mis-specification in the construction of the denominator of (6), as shown.

Note that we evaluate the forecast densities separately for each horizon, and the timing of the forecasts (explained in the next section) is such that the forecasts are non-overlapping. They are non-overlapping in the sense that the realization for the previous year's Q1 survey forecast (say) is known before this year's Q1 survey forecast is made. Hence for correctly-specified forecasts, we would expect to be able to set $\gamma_i = 0$ for $i > 0$ in estimating the asymptotic variance of $\overline{d}_h$. However, for dynamically mis-specified densities we may

choose to use an autocorrelation-consistent estimator of the variance. [5]

3 Forecast data description

We use the quarterly US Survey of Professional Forecasters (SPF) respondents' probability distributions for inflation and output growth. The SPF began as the NBER-ASA survey in 1968:4 and runs to the present day: see Croushore (1993). It has been extensively used for academic research into the nature of expectations formation: see https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/academic-bibliography.

We use the forecast distributions (provided in the form of histograms) from the 130 quarterly surveys from 1981:Q3 to 2013:Q4, inclusive, as our primary source of data.[6] Over this period, the survey provides respondents' histograms for output growth and inflation, in terms of the percentage change in the year of the survey relative to the previous year.[7] For surveys which take place in the first quarters of the year, the latest available GDP and GDP deflator data are advance estimates for the fourth quarter of the previous year, so the forecast horizon is effectively 4-quarters ahead. The next quarter - the second quarter of the year - the target is again the annual calendar year growth rate of the current year relative to the previous year, but now the first quarter advance estimates of the national accounts data are known, and the forecast horizon is 3 quarters. The shortest horizon therefore occurs for fourth quarter surveys, where there is data on all but the fourth quarter of the year.

We suppose that the forecasters are implicitly targeting an early vintage release. That is, the distributions are compared to the advance estimates of calendar-year output growth and inflation released in Q1 of the year following the year being forecast. [8] To be explicit, the actual calendar-year percentage growth rate (for GDP or its deflator) for the 1981:Q3 survey is given by:

$$100 \times \left(\frac{Y^{82:1}_{81:1} + Y^{82:1}_{81:2} + Y^{82:1}_{81:3} + Y^{82:1}_{81:4}}{Y^{82:1}_{80:1} + Y^{82:1}_{80:2} + Y^{82:1}_{80:3} + Y^{82:1}_{80:4}} - 1 \right) \tag{7}$$

where Y refers to the (non-logged) value of the variable in the quarter given by the subscript, from the data vintage given by the superscript. The superscript here makes clear that all the data come from the 1982:Q1 vintage. The numerator is the sum of the quarterly values of the variable in the current year, and the denominator is the same for the previous year, where the current and previous year are relative to

[5] When we do so, we use the standard approach and estimate ξ^2 by $\hat{\xi}^2$ where $\hat{\xi}^2 = \hat{\gamma}_0 + 2\sum_{j=1}^{p}\left(\frac{p-j}{p}\right)\hat{\gamma}_j$, where $\hat{\gamma}_0 = \frac{1}{n}\sum_{t=1}^{n}\left(d_t - \overline{d}\right)^2$, $\hat{\gamma}_j = \frac{1}{n}\sum_{t=j+1}^{n}\left(d_t - \overline{d}\right)\left(d_{t-j} - \overline{d}_j\right)$, and $d_t \equiv d_{t|t-h}$, $\overline{d} = \frac{1}{n}\sum_{t=1}^{n} d_t$, $\overline{d_j} = \frac{1}{n-j}\sum_{t=j+1}^{n} d_{t-j}$.

[6] The data were downloaded in December 2015, from http://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/

[7] Also provided are histograms for the year-on-year calendar growth rates for the next year relative to the year of the survey, but we do not analyze these.

[8] These are taken from the quarterly Real Time Data Set for Macroeconomists (RTDSM) maintained by the Federal Reserve Bank of Philadelphia: see Croushore and Stark (2001). This consists of a data set for each quarter that contains only those data that would have been available at a given reference date: subsequent revisions, base-year and other definitional changes that occurred after the reference date are omitted.

the year the survey is held in. In order to forecast the calendar-year growth rate given by (7), the survey respondent will need forecasts of the remaining quarters of the current year, $Y_{81:3}$ and $Y_{81:4}$. Estimates of the earlier two quarters of 1981 are available, but these are the 1981:Q3 vintage estimates, and are subject to revision relative to the 1982:Q1 vintage values used to generate the actual growth rate given by (7). (This also applies to the estimates of the 1980 calendar year values.) One would expect the effects of the revisions to the 'known-quarters' ($Y_{81:2}$, $Y_{81:1}$, $Y_{80:4}$, ..., $Y_{80:1}$) to be of secondary importance relative to the errors in forecasting the future quarters.

The target remains the same for the 1981:Q4 survey forecasts, and now the only unknown $Y_{81:4}$, so intuitively there is less uncertainty about the calendar-year growth rates than in the previous quarter.[9]

The next quarter, 1982:Q1, the target switches to the 1982 calendar year growth, and remains so for the 1982 quarter surveys, and so on.

The sample begins with the 1981:Q3 survey, because prior surveys asked for probability distributions for nominal (as opposed to real) output. In fact there are a number of other complicating factors, to do with changes in variable definitions (GNP to GDP), base year changes, the size and locations of the histogram bins, the number of respondents, and the years to which the forecasts refer.[10]

The survey also provides point forecasts of the calendar-year GDP and the GDP deflator, which match the histograms in that they are fixed-event (see, e.g., Nordhaus (1987)) - forecasts of the same target (here, the annual calendar year growth rates) made at a number of forecast origins (here, the 4 surveys of the year). Rolling-event forecasts of GDP and its deflator are also made. These are forecasts of the values of the variables in the current and each of the next four quarters. These forecasts will be used in conjunction with the real-time data set to construct the benchmark densities, as described in section 4. We are able to use the rolling-event (or 'fixed-horizon') point forecasts from surveys prior to 1981:Q3. Although the output forecasts prior to 1981:Q3 originally referred to nominal GDP (GNP), forecasts of real GDP (GNP) have been constructed by the survey administrator using the forecasts of the deflator.

The probability assessments are reported as histograms, which provide an incomplete estimate of the densities. We fit normal distributions to the histograms (see, e.g., Giordani and Söderlind (2003, p. 1044) and Boero, Smith and Wallis (2015)) when there are 3 or more bins with non-zero probability attached, and otherwise we fit triangular distributions, in precisely the same way as explained in Engelberg *et al.* (2009, see p.37-8). From these distributions we obtain estimates of z and log scores. In order to do so, we follow much of the literature and assume the open-ended exterior bins are in fact closed, by assuming they have the same

[9] This is formalized by Manzan (2016) within a Bayesian framework in which an individual updates her/his prior density as new information becomes available.

[10] See, for example Diebold, Tay and Wallis (1999), Clements (2010) for a discussion of some of these aspects, as well as the online documentation provided by the SPF at https://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters. Many of these potential complicating factors can be readily dealt with, e.g., we omit the 1985:Q1 and 1986:Q1 since there is some doubt as to the years to which the survey histogram questions refer. Others, such as the changing composition of the 'panel' of forecasters potentially have more wide-reaching effects. For example, we assume that data are missing 'at random' so that the sample is representative of the population. That is, failure by an individual to respond to a given survey occurs for reasons unrelated to the issues of interest in our study. López-Pérez (2015) is one of the few studies to consider whether the decision to participate is related to (say) perceived uncertainty about the outlook.

width as the interior bins. This assumption is innocuous when most if not all of the probability is assigned to inner bins, but matters when the location of the bins lags developments in the economy, resulting in a pile up of probability in an open bin, and effectively a truncated distribution. This only appears to have been a potential problem for output growth in the 2009:Q1 and Q2 surveys for the SPF. In 2009:Q1 the aggregate histogram assigned a 34% chance to (2009-calendar-year) growth being less than -2. For the 2009:Q2 survey, a new lower bin was added, ' < -3', and the aggregate histogram indicated a 24% chance of this event. But as noted by Manzan (2016), there are no other quarters where a large average probability is assigned to an open interval, so that any distortionary effects are likely to be minor (and have no effect on the Q3 and Q4 survey forecasts). In contrast, Manzan (2016) shows that the problem is more acute for the European Central Bank SPF, and suggests fitting 'artificial' triangular distributions, based on the assumption that an individual's point prediction can be regarded as an estimate of the mode of his/her underlying distribution.

Finally, as explained in this section, the SPF histogram forecasts are fixed-event, in that there are a number of forecasts of the same target of different horizons, e.g., forecasts of the 2010 calendar-year growth rate made in each of the four quarters of 2010. Then the 2011:Q1 survey targets the 2011 calendar-year growth rate, as do the remaining quarters of 2011, and so on. This has the major advantage for our purposes of delivering series of annual forecasts of horizons of 1, 2, 3 and 4 quarters ahead. It allows a study of the term structure, and the behaviour of the histograms as the horizon shortens. [11] The European Central Bank SPF provides both fixed-horizon 1-year ahead and 2-year ahead annual growth rate histogram forecasts, as well as fixed-event histogram forecasts, but many of the studies of the ECB-SPF use the medium or long-term fixed horizon forecasts (e.g., Abel, Rich, Song and Tracy (2016) and Kenny, Kostka and Masera (2014, 2015)) and are silent about the short-horizon properties of the forecasts.

4 The benchmark density forecasts.

Our benchmark density forecasts (BMs) are constructed from the historical distributions of past median SPF point prediction forecast errors. We centre the densities on the median point predictions of the annual year-on-year growth rates, in which case the BM densities condition on the location, but not the scale of the distribution. We use the median, as opposed to the mean of the cross-section of point forecasts, because the median is usually taken to be the consensus. We also calculate truly unconditional BMs, as explained below. In either case, the key input is the set of rolling-event forecasts of the quarterly values of horizons up to 4-quarters ahead. These forecasts are used to construct *ex ante*, real-time distributions of forecast errors that are comparable to the errors in forecasting the calendar-year annual growth rates. We require: (*i*) that the forecast horizons match, and (*ii*) that the targets match. We explain how this is achieved by way of an example.

[11] D'Amico and Orphanides (2008) propose a way of constructing approximate 1-year fixed horizon forecasts for the US SPF by weighting together the current calendar-year forecasts and the forecasts of next year's calendar-year growth rate.

Consider the construction of the benchmark density for the 1981:Q3 survey. The histogram forecasts made in response to this survey are 2-quarters ahead (in the sense that the latest data values for output growth or inflation are for 1981:Q2). Hence we need 2-step ahead forecast errors to construct the BMs. They must also be of the annual growth rate relative to the previous year. We calculate 50 2-step ahead annual growth rate forecast errors, using the 1968:Q4 survey to the 1981:1Q1 survey, inclusive. That is, the forecast errors would have been known at the time, so a density based on these errors could in principle have been constructed, and used to forecast. In this sense the benchmark density forecasts are real time.

The first of the 50 forecast errors is given by:

$$\overbrace{100 \times \left(\frac{Y^{81:3}_{81:2} + Y^{81:3}_{81:1} + Y^{81:3}_{80:4} + Y^{81:3}_{80:3}}{Y^{81:3}_{80:2} + Y^{81:3}_{80:1} + Y^{81:3}_{79:4} + Y^{81:3}_{79:3}} - 1 \right)}^{\text{Actual Value}} - \overbrace{100 \times \left(\frac{F^{81:1}_{81:2} + F^{81:1}_{81:1} + Y^{81:1}_{80:4} + Y^{81:1}_{80:3}}{Y^{81:1}_{80:2} + Y^{81:1}_{80:1} + Y^{81:1}_{79:4} + Y^{81:1}_{79:3}} - 1 \right)}^{\text{Forecast Value}} \tag{8}$$

where F^o_q refers to the forecast of target quarter q made at time (forecast origin o), and as in section 3, Y^v_q refers to the actual value of the variable in quarter q taken from data vintage v. The following points should be noted: 1) the actual value is the value of the four quarters through 1981:Q2 as a percentage of the previous four quarters, all taken from the 1981:Q3 vintage of data; 2) 'the' forecast consists of forecasts of two quarters, 1981:Q2 and 1981:Q1, both from the 1981:Q1 survey, and data values for the previous six quarters[12]; 3) the denominators in the Actual and Forecast growth rates refer to the same quarters (i.e., subscripts match), but the superscripts differ - for the Forecast value the actuals in the denominator come from the then available (1981:Q1) vintage.

The second of the fifty forecast errors (used to construct the benchmark density for the 1981:Q3 survey) is as (8) but based on a two-step ahead 'year-on-year' forecast from the 1980:4 survey, that is:

$$\overbrace{100 \times \left(\frac{Y^{81:2}_{81:1} + Y^{81:2}_{80:4} + Y^{81:2}_{80:3} + Y^{81:2}_{80:2}}{Y^{81:2}_{80:1} + Y^{81:2}_{79:4} + Y^{81:2}_{79:3} + Y^{81:2}_{79:2}} - 1 \right)}^{\text{Actual Value}} - \overbrace{100 \times \left(\frac{F^{80:4}_{81:1} + F^{80:4}_{80:4} + Y^{80:4}_{80:3} + Y^{80:4}_{80:2}}{Y^{80:4}_{80:1} + Y^{80:4}_{79:4} + Y^{80:4}_{79:3} + Y^{80:4}_{79:2}} - 1 \right)}^{\text{Forecast Value}} \tag{9}$$

Continuing back in time in this way results in the 50th forecast error being of the four quarters up to and including 1969:Q1, relative to the four quarters through 1968:Q1, all taken from the 1969:Q2 vintage, compared to the forecast of this quantity. This comprises forecasts from the 1968:Q4 survey of 1969:Q1 and 1968:Q4, and values of 1968Q3 and Q2 from the 1968:Q4 data vintage.

To generate a BM to match a fourth quarter survey density (such as 1981:Q4, to make things concrete), the forecast errors need to reflect the fact that there is one fewer quarter to forecast. The first forecast error is constructed as:

[12] The SPF provides 'forecasts' of the previous quarter's value, and these are invariably set equal to the released data, i.e, $F^{81:1}_{80:4} = Y^{81:1}_{80:4}$.

$$\overbrace{100 \times \left(\frac{Y^{81:4}_{81:3} + Y^{81:4}_{81:2} + Y^{81:4}_{81:1} + Y^{81:4}_{80:4}}{Y^{81:4}_{80:3} + Y^{81:4}_{80:2} + Y^{81:4}_{80:1} + Y^{81:4}_{79:4}} - 1 \right)}^{\text{Actual Value}} - \overbrace{100 \times \left(\frac{F^{81:3}_{81:3} + Y^{81:3}_{81:2} + Y^{81:3}_{81:1} + Y^{81:3}_{80:4}}{Y^{81:3}_{80:3} + Y^{81:3}_{80:2} + Y^{81:3}_{80:1} + Y^{81:3}_{79:4}} - 1 \right)}^{\text{Forecast Value}} \tag{10}$$

and the second as:

$$\overbrace{100 \times \left(\frac{Y^{81:3}_{81:2} + Y^{81:3}_{81:1} + Y^{81:3}_{80:4} + Y^{81:3}_{80:3}}{Y^{81:3}_{80:2} + Y^{81:3}_{80:1} + Y^{81:3}_{79:4} + Y^{81:3}_{79:3}} - 1 \right)}^{\text{Actual Value}} - \overbrace{100 \times \left(\frac{F^{81:2}_{81:2} + Y^{81:2}_{81:1} + Y^{81:2}_{80:4} + Y^{81:2}_{80:3}}{Y^{81:2}_{80:2} + Y^{81:2}_{80:1} + Y^{81:2}_{79:4} + Y^{81:2}_{79:3}} - 1 \right)}^{\text{Forecast Value}} \tag{11}$$

and so on.

To generate BM densities to match first quarter survey densities, all the quarters in the numerator of the year-on-year growth rate forecast will need to be replaced by forecast values.

In all relevant respects the forecasts underlying the errors in (8) to (11) mirror the corresponding histogram forecasts, and so the distribution of these forecast errors can be used as a valid benchmark for comparison. We fit normal distributions to these. For example, the 1981:Q3 truly unconditional distribution is assumed to be normal with mean given by the average of the past forecasts, and variance given by the sample variance of the past forecast errors. Because the forecasts used to calculate the forecast errors are overlapping, we use an autocorrelation-correction. Specifically, we set $p = 5$ in the footnote to section 2.2. For the benchmark forecasts that condition on the location, we set the mean to the median SPF point prediction of the rate of growth of 1981 over 1980 made in response to the 1981:Q3 survey, and calculate the variance as for the truly unconditional densities. Fitting a normal facilitates the calculation of z^* as well as the log score for comparison to the SPF histogram-based distributions.[13]

Our choice of benchmark forecasts are motivated in part by Rossi and Sekhposyan (2015). They use the empirical distribution of SPF forecast errors to construct an uncertainty index, based on the percentile of the historical distribution which corresponds to the forecast error for the particular realization: forecast errors in the tails are deemed more uncertain than those away from the tails. Our focus is different, but nevertheless past SPF forecast errors provide distributions against which the SPF conditional distributions can be compared.

5 Empirical Results

5.1 The Aggregate Distributions

Aggregate distributions calculated by equal weighting of the individual respondents' forecast distributions are often the object of the analysis. Equal weighting is known in the literature as the linear opinion pool

[13] We do not need to assume normality to calculate z (and hence z^*): we could simply look at the proportion of the historical errors which are less than the realization. However, when this is 0 or 1, the calculation of z^* is problematic as the inverse normal cdf is not defined.

(see Genest and Zidek (1986)), and such forecasts are reported by the US SPF along with the individual histograms, although there are other ways of combining density forecasts (Hall and Mitchell (2009) provide a review). Denoting individual i's density forecast for Y_t made at time $t-h$ by $p_{Y,i,t-h}(y_t)$, with mean and variance $\mu_{i,t|t-h} = \int_{-\infty}^{\infty} y_t p_{Y,i,t-h}(y_t)\,\partial y_t$ and $\sigma^2_{i,t|t-h} = \int_{-\infty}^{\infty} (y_t - \mu_{it})^2 p_{Y,i,t-h}(y_t)\,\partial y_t$, for $i = 1, \ldots, N$ the aggregate density is: $p_{Y,t-h}(y_t) = \frac{1}{N}\sum_{i=1}^{N} p_{Y,i,t-h}(y_t)$ with mean and variance given by:

$$\begin{aligned} \mu_{t|t-h} &= \tfrac{1}{N}\sum_{i=1}^{N} \mu_{i,t|t-h} \\ \sigma^2_{t|t-h} &= \tfrac{1}{N}\sum_{i=1}^{N} \sigma^2_{i,t|t-h} + \tfrac{1}{N}\sum_{i=1}^{N} \left(\mu_{i,t|t-h} - \mu_{t|t-h}\right)^2. \end{aligned} \tag{12}$$

Hence the mean of the aggregate distribution is the simple average of the means of the individual distributions, whereas the variance is the average of the individual variances plus the second term which measures disagreement between forecasters, and serves to increase the aggregate variance relative to the cross-section average: see Wallis (2005), extending earlier work by e.g., Lahiri, Teigland and Zaporowski (1988).

Use of the aggregate histograms allows unbroken sequences of forecasts across the entire sample, as the average is taken across all respondents (with 'N' varying over t), and the changing composition is ignored. For many individuals there are many non-response surveys, generally due to late joining or leaving the survey, which is obviously exacerbated by the survey's long historical duration (relative to similar surveys, such as those run by the ECB and Bank of England, for example). The aggregate histograms are often regarded as a summary measure of the information in the survey, in much the same way as the median point prediction is often taken as *the* survey point prediction, and used in comparisons with model forecasts, for example. But whereas the average point prediction is a 'good' summary measure (see, e.g., Clemen (1989), Aiolfi, Capistrán and Timmermann (2011) and Manski (2011) as examples of a very large literature), the same may not be true of density combination because of the inflation of the variance of the aggregate histogram (relative to, say, a randomly-selected individual histogram). We address this issue below by using a measure of the average variance which abstracts from the disagreement term in (12).

The aggregate histograms always allocate non-zero probability to more than 2 bins, so that the normal distribution is fit to all the histograms, and is used to calculate z^* and the log scores.[14]

The results are recorded in table 1. Consider first the SPF densities (not 'corrected' for the effects of disagreement) - these are the first rows for each forecast quarter headed simply by 'SPF'. For the Q1 surveys (corresponding to a 4-quarter ahead horizon) the output growth forecasts are not rejected using any of the three Berkowitz tests, but the output growth densities are rejected for the three other survey quarters (so for forecast horizons of 3, 2 and 1 quarters ahead). The SPF inflation forecast densities are rejected for all survey quarters. As noted above, a possible explanation of the rejection of the aggregate SPF forecasts is

[14] When, as here, z is calculated after fitting a Gaussian distribution to the histogram, taking the inverse standard normal cdf of z to give z^* results in $z^* = (y - \hat{\mu})/\hat{\sigma}$, where y is the realization, and $\hat{\mu}$ and $\hat{\sigma}^2$ are the mean and variance of the fitted distribution.

that disagreement inflates the variances of the aggregate histograms, leading to the histograms being 'too dispersed' given the realizations, and the z^* variables showing too little dispersion: this is consistent with the estimates of the variances of the AR(1) model fit to $\{z_t^*\}$ reported in table 1. One way of dealing with this concern is to simply calculate the variance of the aggregate histogram as the cross-sectional average of the individuals' histogram variances, setting the second term in (12) to zero.[15] The results of doing so are recorded in the rows headed 'SPF (no disag.)'. There is no change regarding inference based on the Berkowitz tests at the 5% level. There are some minor changes, mainly at the longer horizons. For example, the Q1 inflation densities are no longer rejected at the 1% level on the 2 and 3-degree of freedom Berkowitz tests (but they would be at the 2% level).

The results of the raw moment tests of Knüppel (2015) also suggest the SPF densities are inadequate. The test that the z^* are standard normal based on the first four raw moments (column headed 'First 4') rejects the Q1 survey output growth forecasts: of all the SPF forecasts, only the Q1 output growth forecasts corrected-for-disagreement are not rejected.

Next, we consider the adequacy of the two sets of Benchmark forecasts. Except at the longest-horizon first-quarter surveys, the adequacy of these forecasts is resoundingly rejected. For inflation, the first quarter Benchmark forecasts (which condition on location, denoted simply 'BM' in the table) are not formally rejected on any of the 3 Berkowitz tests at the 5% level, but the unconditional Benchmark ('BM (truly uncond.)') is rejected for inflation. For output growth, and the first quarter surveys, we find the reverse.

Having found all densities (SPF and the Benchmarks) mis-specified on the Berkowitz tests for all but the longest horizon forecasts, what can we say about the relative comparison of the SPF against BM (on log score)? The final column of the table reports comparisons of the forecast densities against BM based on KLIC (or log scores) using equation (6). The entries in the tables are constructed such that a value less than 0.05 leads to a rejection of equal accuracy in favour of the BM at the 5% level (one-sided test), and an entry greater than 0.95 rejects in favour of the alternative (SPF, SPF (no disag.) or BM (truly uncond.)) being more accurate (at the 5% level in a one-sided test). The aggregate SPF are more accurate at the two longer horizons (the first two quarters of the year) for output growth, no better or worse for third quarter surveys, and less accurate at fourth quarter surveys. For inflation, the SPF are no better at the longest horizon, but worse at the other three shorter horizons.

Moreover, the *comparison* of the two sets of Benchmark densities for both variables and for all survey quarters clearly favours the BM forecasts which condition on location.[16] For both variables and all survey

[15] In fact we calculate the standard deviation of the aggregate histogram as the average of the individuals' standard deviations. The results of doing so are indistinguishable from taking the square root of the average of the individuals' variances. Abel *et al.* (2016) calculate the 'average variance' by first estimating the variance of the aggregate histogram, and then subtracting the disagreement term. See also Lahiri and Sheng (2010) and Lahiri, Peng and Sheng (2015).

[16] Including for the first-quarter output growth forecasts, that is, notwithstanding the failure to reject the unconditional Benchmark densities on the Berkowitz tests, and the rejection of those which condition on location. The Berkowitz tests reject in part because the estimated variance of z^* is less than 1 ($\hat{\sigma}^2 = 0.23$ for Q1 output growth forecasts). For the truly unconditional densities, the large forecast errors which result from not conditioning on location result in more extreme values of z^* and an estimated variance close to 1.

quarters we reject in favour of BM against the unconditional BM at the 5% level.

In summary, then, the survey densities fare worse as the forecast horizon shortens. For output growth the survey densities are only rejected at the shortest horizon, while for inflation the survey densities are rejected at all but the longest (4-quarter) horizon.

The results hold irrespective of how disagreement is treated. The reported results set the covariance terms to zero in the calculation of the asymptotic variance of the Diebold-Mariano statistic (6). The results did not change to any significant degree if we used instead an autocorrelation-consistent estimator of the standard error.[17]

5.2 The Individual Distributions

We report results for all individuals who filed more than 15 forecasts of a given horizon. There is no requirement that these are made in consecutive years. We fit normal distributions when histograms have 3 or more non-zero bins, and (isosceles) triangular distributions when either one bin has all the probability mass, or it is distributed across two (adjacent) bins. Triangular distributions result in values of z of 0 or 1 when the realization falls outside of the support: these are arbitrarily set to 0.01 and 0.99, respectively. (Then $z^* = \pm 2.3263$ and is well defined, and an interpretation is that no realized values are viewed as 'impossible' by the SPF respondents). We make the same assumption when we calculate log scores. However this assumption is arbitrary, resulting from the log score not being defined for outcomes with zero forecast probability. Boero, Smith and Wallis (2011) suggest the use of alternative scoring rules for histogram-based probability assessments, as such eventualities are likely to arise when forecasts are presented as histograms (see, also, Kenny *et al.* (2014)). We consider the use of alternative scoring rules arguably better suited to assessing histograms in section 6.4.

Tables 2 and 3 report results for output growth and inflation. We report a two-degree of freedom test of zero-mean and unit-variance, but do not allow an AR(1) as the unrestricted model because there are fewer observations for individuals, and because there are missing values, which would complicate the fitting of an autoregressive model. (Moreover, the results for the aggregate histograms suggest that power to reject the null is more likely to come from testing the mean and variance of z^*, not from testing for autocorrelation). For each individual and horizon, then, we report the number of forecast observations, the p-value of the Berkowitz two-degree of freedom test, the estimates of the mean and variance, and the Diebold-Mariano test of the log scores. The results for individuals are sorted by p-value of the Berkowitz test within each survey quarter (equivalently, forecast horizon).

Consider table 2 for output growth. Except at the longest horizon, the statistical adequacy of the SPF densities is rejected for fewer than one third of respondents at the 5% level. (Specifically, for Q2, 2 of 12; for Q3, 2 of 9; and for Q4, 4 of 13). This suggests the scales of the individual respondents' histograms may be

[17] We used $p = 3$ in the expression given in the footnote in section 2.2.

better calibrated than the scales of the aggregate distributions, with two thirds of individuals reporting 1-quarter ahead (Q4 survey forecasts) which are well calibrated. However, the fewer rejections at the individual level may also reflect lower power due to the smaller sample sizes. Tellingly, the BM forecasts (conditioning on location) are statistically more accurate on log score than the forecasts of each individual SPF respondent at the shortest horizon (i.e., for Q4 surveys). (The entries in the table are constructed as in table 1, such that a value below 0.05 suggests the BM are more accurate at the 5% level (one-sided test), and an entry greater than 0.95 rejects in favour of the SPF being more accurate (at the 5% level in a one-sided test). At the longer horizons, the SPF forecasts are not rejected against the BMs (i.e., for the Q1 and Q2 survey quarters) but instances of rejecting in favour of an SPF respondent are rare. In summary, the comparisons against the BM forecasts suggests that the individual SPF output growth densities are poor at the shortest horizon.

For inflation (table 3) the rejection of the SPF forecasts is less equivocal, in that the SPF forecasts are rejected on the Berkowitz test for 7 out of 8 and 9 out of 12 respondents for Q3 and Q4 forecasts, respectively, and in addition, the forecasts of all these individuals are found to be statistically less accurate than the BM forecasts for most individuals except at the longest horizon, Q1 surveys.

6 Robustness of the results to the assumptions

We consider whether the findings discussed in section 5 are unduly dependent on *i*) the assumptions we have made in order to construct forecast densities from the histograms, and *ii*) the use of log score as the density scoring rule, given that the probability assessments come from histograms.

6.1 The wider histogram bin widths before 1992:Q1.

Prior to 1992, respondents assigned probabilities to intervals of width 2 percentage points. From 1992Q1 onwards a finer gradation was adopted with intervals being halved. The use of wide intervals may give a misleading picture when uncertainty is low, and all the probability is assigned to one interval. In such circumstances, a symmetric triangular distribution (with support on the full interval) results in a variance of 0.0416 when the interval width is one, but (four times as large) at 0.1666 when the interval width is two. Individuals are more likely to assign all the probability to one interval in response to Q4 surveys, because perceived uncertainty will be smaller at the shortest horizon. Hence our approach will place a floor on the variance (and will affect the z and log score calculations) for earlier-period one-bin histograms which may be particularly distortionary prior to 1992.

We tackle this in two ways.

Firstly, we approximate the one-bin histograms prior to 1992 by a symmetric triangular distribution with a base of one (rather than two), located centrally within the interval. There is no way of knowing whether this provides a more accurate representation of the underlying subjective distribution. We simply wish to

assess whether the way we treat the wide single-bin histograms is driving the results. This has no effect on the aggregate results for either variable, because all the aggregate histograms allot probability to 3 or more bins. For individuals the change is most likely to affect the shorter horizon forecasts. For the shortest horizon for output growth there are inconsequential changes. We still reject for 4 of the 13 respondents, and for all respondents reject in favour of the BMs on log score comparisons. (These results are not shown). The results for inflation are also unchanged.

Secondly, we set the beginning of the forecast sample to 1992:Q1. Some authors have regarded the post 1992 period of the SPF as providing cleaner forecast data under the stewardship of the Philadelphia Fed, and it avoids the drop off in participation rates that occurred in the 1980s (see, e.g., Engelberg *et al.* (2009) and Manzan (2016)). Curtailing the sample in this way has the drawback of discarding around a third of the observations, and reducing the number of individual respondents we can separately analyze, but also counters the potentially more insidious effects of the wider bins not picked up by the first strategy. The aggregate histogram results for output growth and inflation are broadly unchanged, apart from the SPF output growth forecasts no longer being rejected for Q2 surveys (i.e., for $h = 3$) at the 5% level. Hence we continue to reject the SPF inflation forecasts being well specified at all horizons, and the SPF output growth forecasts are rejected at the two shorter horizons. The pattern of results by individual is also largely unchanged. For Q4 surveys ($h = 1$) we find evidence against the forecasts being well specified for 5 of 11 and 7 of 10 respondents for output growth and inflation, respectively, while for all these respondents we reject in favour of the BMs being more accurate.

6.2 Centreing the SPF densities on the point predictions.

Engelberg *et al.* (2009) find inconsistencies between the central tendencies and the point predictions for some SPF respondents, and Clements (2010) finds evidence that the point predictions tend to be more accurate in terms of traditional forecast evaluation criteria such as squared error loss. This raises the suspicion that the relatively good performance of the BMs may result from their being centred on the (cross-sectional) median *point* prediction. This turns out not to be the case. Centreing the SPF aggregate histograms on the median point predictions, and the individual histograms on the individuals' point predictions,[18] does not result in a marked improvement in the SPF forecasts. (Results for the aggregate and individual densities are available in a Not For Publication Appendix).

6.3 Evaluating regions of the densities

Notwithstanding the poor performance of the short-horizon SPF forecast densities relative to the unconditional benchmark evaluated on log score, the possibility remains that the SPF densities may fare better for

[18] For the individual respondent histograms with probability assigned to 3 bins or more, a normal density is fit to the histogram, as in the standard approach, and the estimated mean is then replaced by the individual's point prediction. For the one and two bin histograms for which we assume triangular distributions no use is made of the point prediction.

particular regions of the density, such as the 'event' defined by output growth being less than some threshold value, for example, or of inflation being in the vicinity of 2%.[19]

We might view the SPF output growth forecasts much more favourably if it were the case that the densities implied high forecast probabilities of declines in output when output did indeed fall.[20] Or if the inflation densities were accurate for rates of inflation in the region of 2%. Because there were only three calendar years of negative annual growth during our sample, there are insufficient observations of declines in output to reliably compare SPF and benchmarks densities in terms of this event. Instead, we consider slower than normal output growth, defined as annual growth less than $1\frac{1}{2}$%, or alternatively, less than 2%. (These events occurred for 6 and 9 calendar years, respectively, so the number of observations underpinning the comparisons is still small). For inflation, we define the region of interest as being $1\frac{1}{2} - 2\frac{1}{2}$%, or $1 - 3$% (occurring 10 and 22 times, respectively).

The region-specific tests of the densities we report are the KLIC-based comparisons of the conditional likelihood score functions of Diks *et al.* (2011, p.219, eqn (9)). That is, instead of comparing the log scores of the SPF and benchmark densities (as in equations (4) to (6)), the difference between the log scores at time t is replaced by the difference between the conditional likelihood scores. The log score for the density, $ln\left(p_{Y,t-h}\left(y_t\right)\right)$, is replaced by:

$$1\left(\tau_1 < y_t < \tau_2\right) \ln\left(\frac{p_{Y,t-h}\left(y_t\right)}{P_{Y,t-h}\left(\tau_2\right) - P_{Y,t-h}\left(\tau_1\right)}\right)$$

where $P\left(\right)$ is the cumulative distribution function corresponding to the density $p\left(\right)$, and $1\left(\right)$ is the indicator function (equal to 1 when the argument is true, and zero otherwise), and τ_2 and τ_1 are the event-defining thresholds, with $\tau_2 > \tau_1$. For output growth, $\tau_1 = -\infty$, and τ_2 is either $1\frac{1}{2}$% or 2%. For inflation, τ_1 and τ_2 are either $\left\{1\frac{1}{2}, 2\frac{1}{2}\right\}$ or $\{1, 3\}$. Diks *et al.* (2011) also propose a censored likelihood approach, and we could alternatively have used the approach of Gneiting and Ranjan (2011).

Table 4 compares the test results using log score (copied from Table 1) with those based on the conditional likelihood scores for the two events of interest for each variable. By and large, the differences are small for both variables, for both events. We conclude that there is no evidence that SPF respondents are relatively better at forecasting regions of the densities corresponding to lower rates of growth of output, or to rates of inflation in the region of the 2% target.

The similarities between the whole density findings and those for the events of interest also serve as a check of a sub-sample (that is, periods of lower rates of growth, or inflation close to 2%) against the whole sample. Further sub-sample analysis could be undertaken to check the constancy of the findings over time, but the nature of the forecast data is such that we only have a relatively small of observations at each horizon. That is, for a given horizon, we only have one forecast observation for each year of the survey.

[19] The 2% target was formally adopted by the FOMC at its meeting in January 2012, and was for the price index for personal consumption expenditures. See, e.g., https://www.federalreserve.gov/faqs/money_12848.htm.

[20] Professional forecasters are not renowned for their ability to predict recessions: see, e.g., Rudebusch and Williams (2009).

6.4 Alternative density scoring rules

Notwithstanding the popularity of the log score in the literature, alternative scoring rules may be preferable for scoring probability forecasts presented in the form of histograms, as noted earlier. In this section we discuss alternative rules and why they may be preferable in our context, and assess the robustness of the results to the scoring rule used.

We consider the quadratic probability score (QPS: Brier (1950)) and the ranked probability score (RPS: Epstein (1969)), defined by:

$$QPS = \frac{1}{N}\sum_{t=1}^{N}\sum_{k=1}^{K}\left(p_t^k - y_t^k\right)^2 \tag{13}$$

and:

$$RPS = \frac{1}{N}\sum_{t=1}^{N}\sum_{k=1}^{K}\left(P_t^k - Y_t^k\right)^2 \tag{14}$$

where there are N forecasts (indexed by t), and for each forecast a probability is assumed to be assigned to each of the K bins (indexed by the superscript k), denoted by p_t^k. y_t^k takes the value 1 when the actual value is in bin k, and zero otherwise. In the definition of RPS, P_t^k is the cumulative probability (i.e., $P_t^k = \sum_{s=1}^{k} p_t^s$), and similarly Y_t^k cumulates y_t^k: $Y_t^k = 1$ for all $k \geq s$, where bin s contains the actual value.

Being based on cumulative distributions, RPS will penalize less severely density forecasts with probability close to the bin containing the actual, relative to QPS. For QPS, a given probability outside the bin in which the actual falls has the same cost regardless of how near or far it is from the outcome-bin. [21]

The notation used in (13) and (14) is deliberately simplified, as we have suppressed the forecast horizon, the individual forecaster (or aggregate), and the fact that the number of bins K changes over time. As described above, QPS and RPS are expressed as loss functions, but in terms of the comparisons reported in the tables are calculated so as to be comparable to log score.

So, QPS and RPS are calculated directly from the SPF histograms using (13) and (14). For the benchmark forecasts, we use the normal distributions (assuming the mean is equal to the median of the cross-section of survey point predictions) that we have estimated from the historical forecast errors, and from these we estimate histograms, using the same number of bins, and location of bins, as underpins the matching SPF forecasts at that point in time. [22]

Table 5 compares the results of testing the SPF histograms against the benchmarks using a Diebold-Mariano test for each of the three scores (the information for the log score repeats that in table 1 and is

[21] In addition, there is a further difference between QPS and RPS given the way we treat the SPF definitions of the bin locations. Suppose the histogram has 100% probability in the SPF interval 6 to 7.9, and obviously 0% in all other intervals, such as the lower adjacent interval defined as 4 to 5.9. The realization is 5.9839. We assume the bins are defined as $[4, 6]$ and $[6, 8]$, and since the realization is in the range (5.95 to 6.05), we assume the actual value is 6 and straddles the two bins. We set $y_t^k = \frac{1}{2}$ for these two bins. For QPS, then, the relevant p_t^k are $0, 1$ and the y_t^k are $\frac{1}{2}$, $\frac{1}{2}$, so QPS is 0.5. For RPS, the P_t^k values are $0, 1$ (as for p_t^k), but Y_t^k is $\frac{1}{2}, 1$, and RPS is 0.25.

[22] In terms of the example in the previous footnote, we evaluate the cdf for the given normal distribution at 8 and 6, and the difference is the probability attached to the bin $[6, 8]$, and so on.

shown for ease of comparison). Note that for the aggregate SPF histograms probability is assigned to 3 or more bins for both variables for all time periods, and so normal distributions are fitted to the histograms, and the issue of outcomes being assigned a non-zero probability can not arise. The broad picture for the aggregate histograms is not changed by using QPS or RPS in place of log score, in the sense that the shortest-horizon (Q4 survey) SPF forecasts are rejected against the Benchmark forecasts at the 5% level. However, a more detailed look suggests there are differences - RPS does not reject the Benchmark forecasts at the longer horizons for output growth, and the Q4 RPS statistic p-value is now 0.01. For inflation, there is no evidence against the SPF forecasts on QPS or RPS at the longest horizon.

The use of QPS and RPS cast the individual SPF forecasters in a more favourable light in terms of output growth, compared to log score: see table 6. For the Q1 and Q2 surveys, few if any SPF individuals' forecasts are rejected (at the 5% level). Whereas on log score nearly a half of the forecasters, and all the forecasters, are rejected in terms of their Q3 and Q4 forecasts, respectively, the proportions rejected on QPS and RPS are markedly lower. Nevertheless, around 40 to 50% of forecasters density forecasts are rejected at Q4, the shortest horizon.[23] For inflation (table 7) the use QPS and RPS suggest around two-thirds of respondents' Q3 forecasts can be rejected, as opposed to all respondents using log score, but at Q4 QPS and RPS are largely in line with log score.

7 Correcting the SPF aggregate density forecasts

Given that our findings appear relatively robust, we consider whether the SPF forecasts can be improved with simple mechanical corrections. Such corrections are commonplace in the point prediction forecasting literature, and are sometimes viewed as a way of 'fixing' a model's forecasts for mis-specification resulting from structural change (see, e.g., Castle, Clements and Hendry (2015)). It is possible to (re-)calibrate future forecast densities for the apparent mis-specification of past densities (for which realizations are available): see, e.g., Dawid (1984), Kling and Bessler (1989) and Diebold, Hahn and Tay (1999). However, given the relatively small number of forecast densities of a given horizon, we consider whether a simple scaling of the SPF aggregate densities would improve their accuracy on log score. Based on the first 15 forecast densities (that is, the densities of 1982 to 1996 for the Q1 and Q2 surveys, or 1981 to 1995 for the Q3 and Q4 surveys), we calculate a horizon-specific scale factor that maximizes log score over this in-sample period,[24] and then

[23] The differences between log score, and QPS and LPS, might be expected to be greatest for the shortest-horizon Q4 survey forecasts. Respondents will typically assign probability to fewer bins, reflecting the lower level of uncertainty about the future, and the problems of 'zero-probability' outcomes resulting from fitting triangular distributions will be more acute.

[24] Given a set of normal forecast densities defined by $\left\{\mu_t, \sigma_t^2\right\}_{t=1}^N$ and realizations $\{x_t\}_{t=1}^N$, the log score is defined by:

$$\sum_{t=1}^N \ln p\left(x_t; \mu_t, \sigma_t^2\right) = -\sum_{t=1}^N \ln \sigma_t - \frac{1}{2}\sum_{t=1}^N \ln 2\pi - \sum_{t=1}^N \frac{\left(x_t - \mu_t\right)^2}{2\sigma_t^2}$$

Choosing λ to maximize the log score over 1 to N, where $\hat{\sigma}_t^2 = \lambda \sigma_t^2$:

$$\sum_{t=1}^N \ln p\left(x_t; \mu_t, \hat{\sigma}_t^2\right) = -\frac{1}{2} N \ln \lambda - \sum_{t=1}^N \ln \sigma_t - \frac{1}{2} N \ln 2\pi - \frac{1}{2\lambda}\sum_{t=1}^N \frac{\left(x_t - \mu_t\right)^2}{\sigma_t^2}$$

apply these factors to the variances of the remaining, out-of-sample density forecasts.

Success would require that the variances of the SPF densities systematically over (or under) estimate the uncertainty over the in-sample period, and that the same remains true of the out-of-sample period. Table 8 records the results for the out-of-sample forecast densities 1996(1997) to 2013.

The first three rows of each panel give the average log scores for the out-of-sample period for the SPF densities, for the SPF densities with the variances scaled by the in-sample estimate of λ, denoted SPF$_c$, and for the BM densities. For the SPF densities we remove the impact of disagreement, and the BM densities are centred on the median point predictions. The DM test results of the (uncorrected) SPF against the BM for the out-of-sample period are similar to the full-sample results in table 1: for output growth the SPF densities are rejected against the BM for Q4 surveys (shortest horizon), and for inflation the SPF densities are less accurate for Q2 to Q4 surveys. Hence the truncation of the assessment period to allow an in-sample period to estimate λ does not affect the findings regarding the SPF aggregate densities.

Table 8 shows marked improvements on log score from scaling-down the variances by the fixed in-sample factors, in all cases except the Q1 output growth forecasts. The improvements in log score result in the Q3 output growth forecasts now being superior to the BMs, and the Q2 SPF inflation forecasts not being rejected as less accurate than the BMs. Not withstanding the improvements in log score at the shortest horizon (Q4 survey quarter) forecasts, these densities are still rejected.

The estimates of λ suggest that the optimal in-sample variances are around a half of the survey variances (for all but the Q1 output growth densities). Applying these adjustments out-of-sample is clearly beneficial, and in some cases changes the inference concerning the relative accuracy of the SPF and BM densities, as noted. That corrections based on past performance up to the mid 1990s yields improvements on average over the remainder of the 1990's and the period up to 2013 suggests systematic failings in the SPF forecasts.

8 Conclusions

Other papers have considered the aggregate US SPF histograms, see, e.g., Diebold *et al.* (1999) and Rossi and Sekhposyan (2013). The novelty of the current contribution is the investigation of the term structure of the aggregate densities and those of individual survey respondents, and the comparison to the benchmark forecasts. The BM densities with means set equal to the median point prediction are designed such that the SPF densities would be superior were the respondents able to gauge the degree of uncertainty characterizing the macro-outlook at the time the forecasts are made.

Our findings suggest that the aggregate and individual forecast densities tend to be too dispersed at the shorter forecast horizons, such as one quarter and two quarters ahead, consistent with the findings of Clements (2014) in his study of *ex ante* and *ex post* forecast uncertainty. Because of the excess dispersion of the survey densities at short horizons, the expected dominance of the SPF densities over the unconditional-

results in $\hat{\lambda} = \frac{1}{N}\sum_{t=1}^{N}\left(\left(x_t - \mu_t\right)/\sigma_t\right)^2$.

variance benchmark distributions at short horizons does not materialize. The benchmark densities are rejected at short horizons on tests of correct specification, as expected, given that the dispersions of these distributions are not conditioned on developments at the time the forecasts are made. However, the excess dispersion on the aggregate SPF densities renders these densities less accurate than the benchmarks on log score comparisons.

The rejection of the short-horizon density forecasts of output growth and inflation occurs for the aggregate densities irrespective of whether an allowance is made for disagreement and irrespective of whether we use log score, QPS or RPS as the scoring rule.

At the individual level, the use of QPS or RPS tends to cast the survey expectations in a more favourable light, compared to when log score is used. Even so, the output growth forecasts of two fifths of the respondents, and the inflation forecasts of four fifths of the respondents, are rejected against the benchmark forecasts for the one-quarter ahead horizon.

Of course the forecasters' subjective probability assessments at short horizons may well be driven by motives other than maximizing accuracy as judged by log score. The respondents' loss functions may be such that the respondents are incentivized to ensure that realized outcomes fall well within the likely range of outcomes implied by their probability assessments, for example. Be that as it may, the excess dispersion of the aggregate histograms at short horizons is sufficiently large and persistent that the application of correction factors (based on a training sample of forecasts and realizations) to out-of-sample probability assessments leads to marked improvements in their (log score) accuracy.

Although we have considered a single macro-survey, it is the longest and probably most used in terms of the probability assessments it provides. Our findings question the reliability of the short-horizon survey densities.

References

Abel, J., Rich, R., Song, J., and Tracy, J. (2016). The Measurement and Behavior of Uncertainty: Evidence from the ECB Survey of Professional Forecasters. *Journal of Applied Econometrics*, *31*(3), 533–550.

Aiolfi, M., Capistrán, C., and Timmermann, A. (2011). Forecast combinations, chapter 11. In Clements, M. P., and Hendry, D. F. (eds.), *The Oxford Handbook of Economic Forecasting*, pp. 355–388: Oxford University Press.

Amisano, G., and Giacomini, R. (2007). Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business & Economic Statistics*, **25**, 177–190.

Andrade, P., and Le Bihan, H. (2013). Inattentive professional forecasters. *Journal of Monetary Economics*, *60*(8), 967–982.

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

Bai, J., and Ng, S. (2005). Tests for Skewness, Kurtosis, and Normality for Time Series Data. *Journal of Business & Economic Statistics*, **23**, 49–60.

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

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

Boero, G., Smith, J., and Wallis, K. F. (2015). The measurement and characteristics of professional forecasters' uncertainty. *Journal of Applied Econometrics*, **30(7)**, 1013–1234.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **75**, 1–3.

Castle, J. L., Clements, M. P., and Hendry, D. F. (2015). Robust approaches to forecasting. *International Journal of Forecasting*, **31**, 99–112.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, **5**, 559–583. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar.

Clements, M. P. (2004). Evaluating the Bank of England density forecasts of inflation. *Economic Journal*, **114**, 844 – 866.

Clements, M. P. (2010). Explanations of the Inconsistencies in Survey Respondents Forecasts. *European Economic Review*, **54(4)**, 536–549.

Clements, M. P. (2014). Forecast Uncertainty - Ex Ante and Ex Post: US Inflation and Output Growth. *Journal of Business & Economic Statistics*, **32(2)**, 206–216. DOI: 10.1080/07350015.2013.859618.

Clements, M. P. (2015). Are professional macroeconomic forecasters able to do better than forecasting trends?. *Journal of Money, Credit and Banking*, **47,2-3**, 349–381. DOI: 10.1111/jmcb.12179.

Clements, M. P., and Galvão, A. B. (2017). Model and survey estimates of the term structure of US macroeconomic uncertainty. *International Journal of Forecasting*, **33**, 591 – 604. DOI: 10.1016/j.ijforecast.2017.01.004.

Coibion, O., and Gorodnichenko, Y. (2012). What can survey forecasts tell us about information rigidities?. *Journal of Political Economy*, *120*(1), 116 – 159.

Coibion, O., and Gorodnichenko, Y. (2015). Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. *American Economic Review*, *105*(8), 2644–78.

Croushore, D. (1993). Introducing: The Survey of Professional Forecasters. *Federal Reserve Bank of Philadelphia Business Review, November*, 3–15.

Croushore, D., and Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics*, **105(1)**, 111–130.

D'Amico, S., and Orphanides, A. (2008). Uncertainty and disagreement in economic forecasting. Finance and economics discussion series 2008-56, Board of Governors of the Federal Reserve System (U.S.).

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of The Royal Statistical Society, ser. A*, **147**, 278–292.

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

Diebold, F. X., Hahn, J. Y., and Tay, A. S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: High frequency returns on foreign exchange. *Review of Economics and Statistics*, **81**, 661–673.

Diebold, F. X., and Kilian, L. (2001). Measuring predictability: Theory and macroeconomic applications. *Journal of Applied Econometrics*, **16**, 657–669.

Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **13**, 253–263.

Diebold, F. X., Tay, A. S., and Wallis, K. F. (1999). Evaluating density forecasts of inflation: The Survey of Professional Forecasters. In Engle, R. F., and White, H. (eds.), *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive Granger*, pp. 76–90. Oxford: Oxford University Press.

Diks, C., Panchenko, V., and van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, *163*(2), 215 – 230. http://dx.doi.org/10.1016/j.jeconom.2011.04.001.

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

Engle, R. F., and Manganelli, S. (2004). CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles. *Journal of Business & Economic Statistics*, **22**, 367–381.

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

Genest, C., and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, **1**, 114–148.

Giordani, P., and Söderlind, P. (2003). Inflation forecast uncertainty. *European Economic Review*, **47(6)**, 1037–1059.

Giordani, P., and Söderlind, P. (2006). Is there evidence of pessimism and doubt in subjective distributions? implications for the equity premium puzzle. *Journal of Economic Dynamics & Control*, **30(6)**, 1027–1043.

Gneiting, T., and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal*

of the American Statistical Association, *102*(477), 359–378.

Gneiting, T., and Ranjan, R. (2011). Comparing density forecasts using threshold and quantile weighted proper scoring rules. *Journal of Business and Economic Statistics*, **29**, 411–422.

Granger, C. W. J., and Pesaran, M. H. (2000). A decision-based approach to forecast evaluation. In Chan, W. S., Li, W. K., and Tong, H. (eds.), *Statistics and Finance: An Interface*, pp. 261–278: London: Imperial College Press.

Hall, S. G., and Mitchell, J. (2009). Recent developments in density forecasting. In Mills, T. C., and Patterson, K. (eds.), *Palgrave Handbook of Econometrics, Volume 2: Applied Econometrics*, pp. 199–239: Palgrave MacMillan.

Hirshleifer, D. (2001). Investor psychology and asset pricing. *Journal of Finance*, **56(4)**, 1533–1597.

Kenny, G., Kostka, T., and Masera, F. (2012). How informative are the subjective density forecasts of macroeconomists?. Working paper series no. 1446, European Central Bank.

Kenny, G., Kostka, T., and Masera, F. (2014). How Informative are the Subjective Density Forecasts of Macroeconomists?. *Journal of Forecasting*, *33*(3), 163–185.

Kenny, G., Kostka, T., and Masera, F. (2015). Density characteristics and density forecast performance: a panel analysis. *Empirical Economics*, *48*(3), 1203–1231.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility : likelihood inference and comparison with ARCH models. *Review of Economic Studies*, **81**, 361–393.

Kling, J. L., and Bessler, D. A. (1989). Calibration-based predictive distributions: An application of prequential analysis to interest rates, money, prices and output. *Journal of Business*, **62**, 477–499.

Knüppel, M. (2015). Evaluating the Calibration of Multi-Step-Ahead Density Forecasts Using Raw Moments. *Journal of Business & Economic Statistics*, *33*(2), 270–281.

Knüppel, M., and Schultefrankenfeld, G. (2012). How Informative Are Central Bank Assessments of Macroeconomic Risks?. *International Journal of Central Banking*, *8*(3), 87–139.

Kullback, L., and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Sciences*, **22**, 79–86.

Lahiri, K., and Sheng, X. (2010). Measuring forecast uncertainty by disagreement:the missing link. *Journal of Applied Econometrics*, **25**, 514–538.

Lahiri, K., Teigland, C., and Zaporowski, M. (1988). Interest rates and the subjective probability distribution of inflation forecasts. *Journal of Money, Credit and Banking*, **20(2)**, 233–248.

Lahiri, K., Peng, H., and Sheng, X. (2015). Measuring Uncertainty of a Combined Forecast and Some Tests for Forecaster Heterogeneity. Cesifo working paper series 5468, CESifo Group Munich.

Lee, T.-H., Bao, Y., and Saltoglu, B. (2007). Comparing density forecast models. *Journal of Forecasting*, *26*(3), 203–225.

López-Pérez, V. (2015). Does uncertainty affect participation in the European Central Bank's Survey of Professional Forecasters?. Working paper series no. 1807, European Central Bank.

Malmendier, U., and Taylor, T. (2015). On the verges of overconfidence. *Journal of Economic Perspectives*, **29**, 3–8.

Manski, C. F. (2011). Interpreting and combining heterogeneous survey forecasts. In Clements, M. P., and Hendry, D. F. (eds.), *Oxford Handbook of Economic Forecasting, Chapter 16*, pp. 457–472. Oxford: Oxford University Press.

Manzan, S. (2016). Are professional forecasters bayesian. mimeo, Baruch College, CUNY, New York.

Mitchell, J., and Hall, S. G. (2005). Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR 'Fan' Charts of Inflation. *Oxford Bulletin of Economics and Statistics*, *67*(s1), 995–1033.

Nordhaus, W. D. (1987). Forecasting efficiency: Concepts and applications. *Review of Economics and Statistics*, **69**, 667–674.

Pesaran, M. H., and Weale, M. (2006). Survey expectations. In Elliott, G., Granger, C., and Timmermann, A. (eds.), *Handbook of Economic Forecasting, Volume 1. Handbook of Economics 24*, pp. 715–776: Elsevier, Horth-Holland.

Rabin, M. (1998). Psychology and economics. *Journal of Economic Literature*, **36(1)**, 11–46.

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

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, **23**, 470–472.

Rossi, B., and Sekhposyan, T. (2013). Conditional predictive density evaluation in the presence of instabilities. *Journal of Econometrics*, *177*(2), 199–212.

Rossi, B., and Sekhposyan, T. (2015). Macroeconomic Uncertainty Indices Based on Nowcast and Forecast Error Distributions. *American Economic Review*, *105*(5), 650–55.

Rudebusch, G. D., and Williams, J. C. (2009). Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve. *Journal of Business & Economic Statistics*, *27*(4), 492–503.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika*, **81**, 115–131.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations*. London: W. Strahan & T. Cadell.

Wallis, K. F. (2005). Combining Density and Interval forecasts: A Modest Proposal. *Oxford Bulletin of Economics and Statistics*, **67(s1)**, 983–994.

Winkler, R. L. (1967). The quantification of judgement: Some methodological suggestions. *Journal of the*

American Statistical Associaton, **62**, 1105–1120.

Wright, J. H. (2013). Evaluating Real-Time VAR forecasts with an informative democratic prior. *Journal of Applied Econometrics*, **28**, 762–776. DOI: 10.1002/jae.2268.

Zarnowitz, V., and Lambros, L. A. (1987). Consensus and uncertainty in economic prediction. *Journal of Political Economy*, **95(3)**, 591–621.

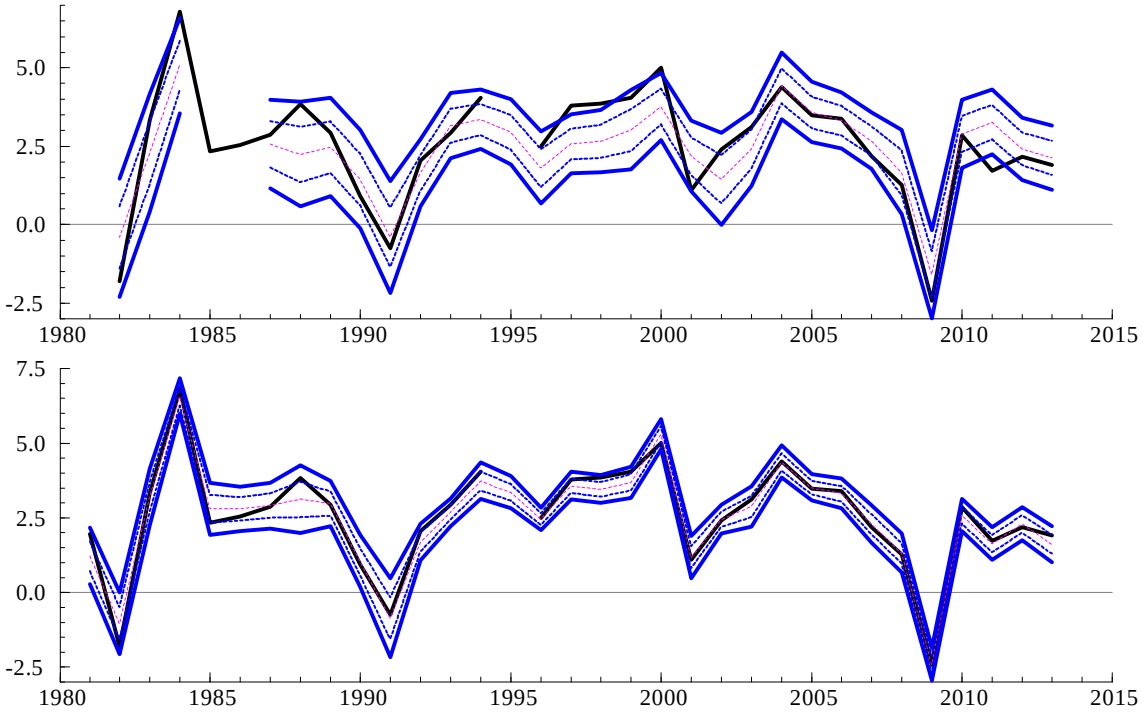

Figure 1: The figure displays the output growth forecasts made in the first quarters of the year (top panel) of the current year-on-year growth rate, and the forecasts made in the fourth quarters of the year (bottom panel). The figures show the 90th and 10 percentiles as solid lines, the 75 and 25th as dotted lines with the median equidistant between these two. The other solid line in each panel are the first-release estimates of the actual growth rates.

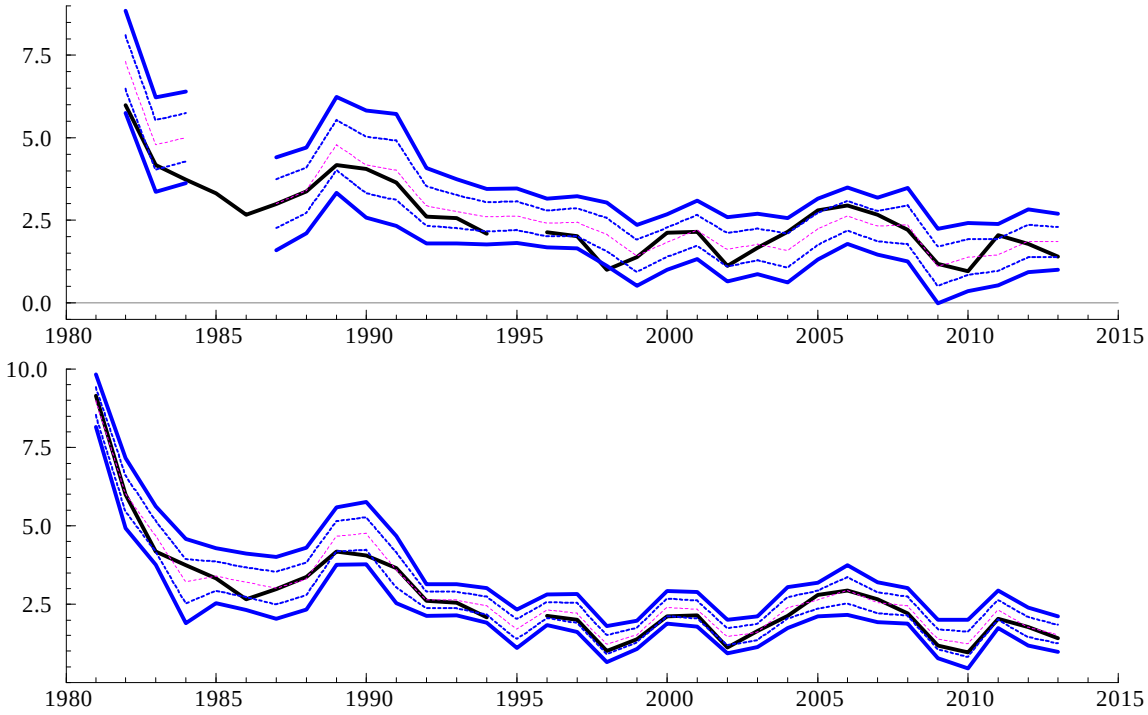

Figure 2: The figure displays the inflation forecasts made in the first quarters of the year (top panel) of the current year-on-year inflation rate, and the forecasts made in the fourth quarters of the year (bottom panel). The figures show the 90th and 10 percentiles as solid lines, the 75 and 25th as dotted lines with the median equidistant between these two. The other solid line in each panel are the first-release estimates of the inflation rates.

Table 1: Aggregate Density Forecasts: SPF & Benchmarks

		Evaluation based on z^*						Raw moments of z^*		Comparison
		Ind.	Eqn (2)	$(0,1)$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\sigma}^2$	First 2	First 4	(Eqn. 6)
Output Growth										
SPF	1	0.14	0.26	0.40	0.18	0.29	0.74	0.24	0.01	1.00
SPF (no disag.)	1	0.22	0.32	0.37	0.23	0.24	0.97	0.47	0.17	0.99
BM	1	0.23	0.00	0.00	0.14	0.24	0.23	0.00	0.00	-
BM (truly uncond.)	1	0.09	0.34	0.75	0.07	0.29	1.08	0.84	0.43	0.02
SPF	2	0.22	0.01	0.01	0.04	0.23	0.39	0.00	0.00	1.00
SPF (no disag.)	2	0.28	0.02	0.02	0.04	0.20	0.40	0.00	0.00	1.00
BM	2	0.81	0.00	0.00	-0.01	0.04	0.20	0.00	0.00	-
BM (truly uncond.)	2	0.11	0.00	0.00	0.09	0.28	2.60	0.14	0.23	0.01
SPF	3	0.09	0.00	0.00	0.14	0.31	0.24	0.00	0.00	0.11
SPF (no disag.)	3	0.09	0.00	0.00	0.12	0.32	0.27	0.00	0.00	0.19
BM	3	0.57	0.01	0.00	0.10	0.11	0.35	0.00	0.00	-
BM (truly uncond.)	3	0.10	0.00	0.00	-0.01	0.30	7.90	0.03	0.02	0.01
SPF	4	0.86	0.00	0.00	0.13	0.03	0.25	0.00	0.00	0.00
SPF (no disag.)	4	0.86	0.00	0.00	0.15	-0.03	0.30	0.00	0.00	0.00
BM	4	0.87	0.00	0.00	0.14	0.03	0.32	0.00	0.00	-
BM (truly uncond.)	4	0.09	0.00	0.00	-0.12	0.31	43.40	0.02	0.01	0.01
Inflation										
SPF	1	0.04	0.00	0.00	-0.13	0.40	0.27	0.00	0.00	0.50
SPF (no disag.)	1	0.09	0.01	0.02	-0.17	0.33	0.39	0.00	0.00	0.88
BM	1	0.09	0.06	0.10	0.03	0.33	0.45	0.03	0.00	-
BM (truly uncond.)	1	0.00	0.00	0.00	-0.39	0.77	0.99	0.00	0.00	0.00
SPF	2	0.03	0.00	0.00	-0.22	0.40	0.17	0.00	0.00	0.00
SPF (no disag.)	2	0.02	0.00	0.00	-0.24	0.42	0.18	0.00	0.00	0.00
BM	2	0.17	0.00	0.00	-0.12	0.25	0.27	0.00	0.00	-
BM (truly uncond.)	2	0.00	0.00	0.00	-0.56	0.78	2.09	0.00	0.00	0.00
SPF	3	0.67	0.00	0.00	-0.37	-0.08	0.12	0.00	0.00	0.00
SPF (no disag.)	3	0.85	0.00	0.00	-0.38	-0.04	0.15	0.00	0.00	0.00
BM	3	0.31	0.00	0.00	-0.30	-0.18	0.26	0.00	0.00	-
BM (truly uncond.)	3	0.00	0.00	0.00	-1.45	0.69	7.09	0.00	0.00	0.00
SPF	4	0.74	0.00	0.00	-0.34	-0.06	0.14	0.00	0.00	0.00
SPF (no disag.)	4	0.69	0.00	0.00	-0.38	-0.07	0.20	0.00	0.00	0.00
BM	4	0.09	0.00	0.00	-0.26	0.30	0.38	0.00	0.00	-
BM (truly uncond.)	4	0.00	0.00	0.00	-3.26	0.70	32.25	0.00	0.00	0.00

The table records the results of evaluating the densities using tests related to Eqn. (2), and comparisons based on Diebold-Mariano tests of equal predictive ability using the KLIC loss function, Eqn. (6).
The first column denotes the survey quarter, whereby '1' indicates a first quarter of the year survey, and a forecast horizon of 4 quarters, and '4' a fourth quarter survey (and a horizon of 1 quarter). The column headed 'Ind.' is the p-value of a test for independence: in terms of Eqn. (2) the test is based on $-2(L\left(\tilde{\mu},\tilde{\sigma}^2,0\right) - L\left(\hat{\mu},\hat{\sigma}^2,\hat{\rho}\right))$, where $\tilde{\mu},\tilde{\sigma}^2$ denote MLEs with $\rho = 0$ imposed.
The next column is the three-degree of freedom test in Eqn. (2), and the column headed $(0,1)$ tests for zero-mean and unit-variance with a maintained hypothesis of independence. The next 3 columns report the estimates of the unrestricted AR(1).
The two columns under the heading 'Raw Moments' are p-values of the Knüppel (2015) tests that the z^* are standard normal, based on the first two, and first four, moments.
The final column is the p-value of the test of the SPF against the Benchmark densities (that condition on location) using Eqn. (6). We also compare the two sets of Benchmark densities one against the other.

Table 2: Evaluation of Individual Respondents' Output Histograms

			Evaluation based on z^*			Comparison (Eqn. 6)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	Log score
421	1	20	0.63	0.11	1.29	0.74
431	1	18	0.22	0.38	1.26	0.43
446	1	18	0.16	0.23	1.65	0.60
99	1	16	0.04	-0.34	1.96	0.89
428	1	16	0.02	0.60	1.63	0.33
411	1	17	0.02	0.34	2.12	0.23
426	1	21	0.01	0.38	1.91	0.29
484	1	17	0.00	0.79	0.89	0.12
20	1	17	0.00	0.38	2.54	0.13
420	1	16	0.00	0.02	3.05	0.24
433	1	17	0.00	0.53	2.58	0.16
407	1	16	0.00	0.92	3.47	0.14
			.	.	.	0.00
421	2	21	0.99	0.00	0.96	0.30
446	2	20	0.80	0.14	0.92	0.62
411	2	16	0.72	-0.16	0.84	0.30
431	2	18	0.44	0.13	0.67	0.85
99	2	16	0.41	-0.09	0.61	0.97
433	2	18	0.21	0.38	1.26	0.62
426	2	19	0.20	0.20	0.58	0.83
463	2	17	0.20	-0.24	0.57	0.56
20	2	23	0.15	-0.11	1.66	0.26
484	2	16	0.15	0.30	0.55	0.10
407	2	16	0.01	0.08	2.39	0.45
65	2	18	0.01	0.13	2.29	0.22
			.	.	.	0.00
84	3	19	0.58	0.19	1.22	0.55
20	3	21	0.50	-0.26	1.03	0.12
421	3	17	0.40	0.33	0.98	0.05
420	3	20	0.23	0.16	0.58	0.36
407	3	19	0.19	0.38	1.28	0.74
433	3	19	0.11	0.42	0.70	0.94
426	3	21	0.04	0.29	0.47	0.04
446	3	18	0.02	0.38	0.41	0.02
65	3	17	0.01	0.35	2.14	0.04
			.	.	.	0.44
84	4	27	0.70	0.15	1.10	0.00
421	4	21	0.61	0.13	0.77	0.00
446	4	17	0.43	0.31	0.99	0.00
411	4	18	0.26	0.23	0.62	0.00
433	4	18	0.25	0.39	1.04	0.00
20	4	20	0.16	-0.18	1.66	0.00
407	4	19	0.08	0.52	1.02	0.01
472	4	16	0.06	0.29	0.44	0.00
431	4	17	0.06	0.57	1.12	0.01
426	4	20	0.01	0.20	0.36	0.00
99	4	18	0.00	0.01	2.54	0.01
420	4	20	0.00	0.36	0.32	0.00
463	4	17	0.00	0.07	0.14	0.00
			.	.	.	1.00

The table records the results of evaluating the densities of individual respondents using tests related to Eqn. (2), and comparisons based on Diebold-Mariano tests of equal predictive ability using the KLIC loss function, Eqn. (6), to compare each individual's forecasts against the Benchmark (where the Benchmark conditions on location).
'N' is the number of forecast densities by the individual 'id' made in response to 'Qtr' surveys. The column headed $(0,1)$ tests for zero-mean and unit-variance with a maintained hypothesis of independence, and the next 2 columns report the estimates of mean and variance.
The final column are the p-values of the test of an SPF individual against the Benchmark densities using Eqn. (6). We also record the percentage of rejections of the null of equal accuracy on log score (in favour of the BM being more accurate at the 5% level) for each forecast horizon .
We consider all respondents who made more than 15 forecasts of a given horizon.

Table 3: Evaluation of Individual Respondents' Inflation Histograms

			Evaluation based on z^*			Comparison (Eqn. 6)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	Log score
411	1	18	0.44	0.23	0.75	0.23
446	1	18	0.23	-0.16	0.56	0.14
433	1	18	0.12	0.05	1.84	0.13
420	1	16	0.12	-0.20	1.83	0.11
484	1	17	0.11	-0.19	0.47	0.01
426	1	21	0.09	-0.34	1.58	0.12
99	1	16	0.03	-0.56	0.60	0.27
421	1	18	0.02	-0.34	0.39	0.03
431	1	19	0.00	-0.08	0.22	0.55
20	1	16	0.00	-0.28	3.52	0.07
407	1	16	0.00	-0.88	2.50	0.02
			.	.	.	0.27
411	2	16	0.54	-0.21	1.27	0.01
431	2	17	0.36	-0.35	0.95	0.04
99	2	17	0.29	-0.35	1.25	0.01
433	2	18	0.20	-0.31	0.65	0.06
65	2	17	0.09	-0.44	1.50	0.10
446	2	20	0.06	-0.33	0.52	0.00
426	2	19	0.04	-0.56	0.77	0.00
463	2	18	0.02	-0.38	0.41	0.00
407	2	16	0.00	-0.87	1.81	0.04
421	2	19	0.00	-1.06	1.73	0.04
20	2	23	0.00	-1.08	3.17	0.00
			.	.	.	0.82
84	3	20	0.07	-0.43	0.64	0.00
426	3	21	0.02	-0.59	0.71	0.00
420	3	20	0.00	-0.44	0.35	0.00
446	3	18	0.00	-0.36	0.24	0.00
407	3	18	0.00	-0.85	0.55	0.00
433	3	19	0.00	-0.31	0.21	0.00
65	3	17	0.00	-0.49	2.76	0.00
20	3	21	0.00	-1.12	1.44	0.00
			.	.	.	1.00
84	4	28	0.34	-0.19	1.32	0.00
431	4	17	0.22	-0.42	1.03	0.00
411	4	18	0.18	-0.11	0.52	0.00
99	4	18	0.04	-0.53	1.46	0.00
463	4	17	0.02	-0.28	0.36	0.00
20	4	19	0.02	-0.65	1.16	0.00
433	4	17	0.01	-0.66	1.50	0.00
426	4	19	0.01	-0.47	0.40	0.00
446	4	18	0.00	-0.47	0.34	0.00
421	4	20	0.00	-0.74	0.77	0.00
407	4	18	0.00	-0.72	1.63	0.00
420	4	20	0.00	-0.57	0.25	0.00
			.	.	.	1.00

See notes to table 2.

Table 4: Evaluating Density Regions Corresponding to Events of Interest

	Comparison based on:		
	Output growth		
Qtr.	Log Score	Conditional $y < 1.5$	Conditional $y < 2$
1	1.00	0.96	0.97
2	1.00	0.96	0.99
3	0.11	0.05	0.15
4	0.00	0.01	0.00
	Inflation		
Qtr.	Log Score	Conditional $1.5 < y < 2.5$	Conditional $1 < y < 3$
1	0.50	0.58	0.67
2	0.00	0.00	0.00
3	0.00	0.00	0.00
4	0.00	0.01	0.00

The table records the results of evaluating the SPF aggregate histograms against the Benchmark using Diebold-Mariano tests of equal predictive ability for log score and for conditional likelihood scores (for the specified events).
The first column denotes the survey quarter, whereby '1' indicates a first quarter of the year survey, and a forecast horizon of 4 quarters, and '4' a fourth quarter survey (and a horizon of 1 quarter).
The 2nd to 4th columns are the p-value of the test of equal accuracy, constructed such that p-values close to zero favour the Benchmarks, and p-values close to one the SPF histograms.

Table 5: Alternative Density Scoring Rules for the Aggregate Density Forecasts

	Comparison based on:		
Qtr.	Log Score	QPS	RPS
	Output growth		
1	1.00	0.92	0.88
2	1.00	0.99	0.94
3	0.11	0.49	0.34
4	0.00	0.00	0.01
	Inflation		
1	0.05	0.59	0.36
2	0.00	0.00	0.01
3	0.00	0.00	0.00
4	0.00	0.00	0.00

The table records the results of evaluating the SPF aggregate histograms against the Benchmark using Diebold-Mariano tests of equal predictive ability for log score, QPS and RPS.).
The first column denotes the survey quarter, whereby '1' indicates a first quarter of the year survey, and a forecast horizon of 4 quarters, and '4' a fourth quarter survey (and a horizon of 1 quarter).
The 2nd to 4th columns are the p-value of the test of equal accuracy, constructed such that p-values close to zero favour the Benchmarks, and p-values close to one the SPF histograms.

Table 6: Alternative Density Scoring Rules for Individual Respondents' Output Histograms

			Comparison based on:		
id.	Qtr.	N.	Log Score	QPS	RPS
421	1	20	0.74	0.46	0.25
431	1	18	0.43	0.53	0.08
446	1	18	0.60	0.50	0.35
99	1	16	0.89	0.38	0.57
428	1	16	0.33	0.07	0.08
411	1	17	0.23	0.11	0.09
426	1	21	0.29	0.41	0.14
484	1	17	0.12	0.27	0.04
20	1	17	0.13	0.43	0.36
420	1	16	0.24	0.23	0.21
433	1	17	0.16	0.17	0.30
407	1	16	0.14	0.19	0.33
			0.00	0.00	0.08
421	2	21	0.30	0.14	0.08
446	2	20	0.62	0.59	0.15
411	2	16	0.30	0.65	0.14
431	2	18	0.85	0.42	0.30
99	2	16	0.97	0.93	0.79
433	2	18	0.62	0.63	0.45
426	2	19	0.83	0.52	0.12
463	2	17	0.56	0.36	0.15
20	2	23	0.26	0.37	0.11
484	2	16	0.10	0.16	0.03
407	2	16	0.45	0.46	0.56
65	2	18	0.22	0.58	0.36
			0.00	0.00	0.08
84	3	19	0.55	0.13	0.12
20	3	21	0.12	0.47	0.52
421	3	17	0.05	0.06	0.02
420	3	20	0.36	0.38	0.45
407	3	19	0.74	0.34	0.35
433	3	19	0.94	0.81	0.81
426	3	21	0.04	0.07	0.06
446	3	18	0.02	0.01	0.01
65	3	17	0.04	0.07	0.09
			0.44	0.11	0.22
84	4	27	0.00	0.29	0.28
421	4	21	0.00	0.07	0.07
446	4	17	0.00	0.00	0.00
411	4	18	0.00	0.03	0.06
433	4	18	0.00	0.15	0.15
20	4	20	0.00	0.01	0.01
407	4	19	0.01	0.12	0.13
472	4	16	0.00	0.61	0.62
431	4	17	0.01	0.07	0.07
426	4	20	0.00	0.00	0.00
99	4	18	0.01	0.01	0.01
420	4	20	0.00	0.00	0.01
463	4	17	0.00	0.04	0.08
			1.00	0.54	0.38

The table records the results of evaluating the densities of individual respondents on log score, quadratic probability score (QPS), and ranked probability score (RPS), relative to the Benchmark, using the Diebold-Mariano test of equal predictive ability (where the Benchmark conditions on location).
The final 3 columns are the p-values of the test of an SPF individual against the Benchmark densities. We also record the percentage of rejections of the null of equal accuracy (in favour of the BM being more accurate at the 5% level) for each forecast horizon, and for each of the 3 scores.
We consider all respondents who made more than 15 forecasts of a given horizon.

Table 7: Alternative Density Scoring Rules for Individual Respondents' Inflation Histograms

			Comparison based on:		
id.	Qtr.	N.	Log Score	QPS	RPS
411	1	18	0.23	0.53	0.22
446	1	18	0.14	0.36	0.29
433	1	18	0.13	0.31	0.19
420	1	16	0.11	0.23	0.11
484	1	17	0.01	0.07	0.09
426	1	21	0.12	0.04	0.07
99	1	16	0.27	0.16	0.15
421	1	18	0.03	0.12	0.09
431	1	19	0.55	0.52	0.39
20	1	16	0.07	0.05	0.05
407	1	16	0.02	0.04	0.02
			0.27	0.18	0.09
411	2	16	0.01	0.00	0.00
431	2	17	0.04	0.01	0.07
99	2	17	0.01	0.10	0.04
433	2	18	0.06	0.06	0.08
65	2	17	0.10	0.89	0.84
446	2	20	0.00	0.01	0.01
426	2	19	0.00	0.12	0.05
463	2	18	0.00	0.01	0.01
407	2	16	0.04	0.02	0.01
421	2	19	0.04	0.00	0.00
20	2	23	0.00	0.00	0.00
			0.82	0.64	0.73
84	3	20	0.00	0.29	0.27
426	3	21	0.00	0.01	0.03
420	3	20	0.00	0.00	0.00
446	3	18	0.00	0.00	0.00
407	3	18	0.00	0.08	0.08
433	3	19	0.00	0.00	0.01
65	3	17	0.00	0.27	0.26
20	3	21	0.00	0.00	0.02
			1.00	0.63	0.63
84	4	28	0.00	0.07	0.07
431	4	17	0.00	0.02	0.01
411	4	18	0.00	0.01	0.03
99	4	18	0.00	0.03	0.03
463	4	17	0.00	0.02	0.08
20	4	19	0.00	0.01	0.01
433	4	17	0.00	0.06	0.04
426	4	19	0.00	0.01	0.01
446	4	18	0.00	0.02	0.02
421	4	20	0.00	0.00	0.00
407	4	18	0.00	0.02	0.03
420	4	20	0.00	0.00	0.00
			1.00	0.83	0.83

The table records the results of evaluating the densities of individual respondents on log score, quadratic probability score (QPS), and ranked probability score (RPS), relative to the Benchmark, using the Diebold-Mariano test of equal predictive ability (where the Benchmark conditions on location).
The final 3 columns are the p-values of the test of an SPF individual against the Benchmark densities. We also record the percentage of rejections of the null of equal accuracy (in favour of the BM being more accurate at the 5% level) for each forecast horizon, and for each of the 3 scores.
We consider all respondents who made more than 15 forecasts of a given horizon.

Table 8: Effects of in-sample adjustment of SPF aggregate density variances out-of-sample

Survey quarter	Q1	Q2	Q3	Q4
	Output Growth			
SPF	-1.28	-0.74	-0.49	-0.13
SPF_c	-1.29	-0.62	-0.24	0.06
BM	-1.44	-0.95	-0.47	0.43
DM:SPF vs. BM	0.88	1.00	0.40	0.00
DB:SPF_c vs BM	0.87	0.99	1.00	0.00
λ	0.97	0.41	0.46	0.54
	Inflation			
SPF	-0.73	-0.52	-0.38	-0.16
SPF_c	-0.66	-0.32	-0.07	0.09
BM	-0.76	-0.25	0.40	1.03
DM:SPF vs. BM	0.72	0.00	0.00	0.00
DB:SPF_c vs BM	0.78	0.14	0.00	0.00
λ	0.47	0.54	0.40	0.37

The first 3 rows in each panel are the average log scores for the years 1997 - 2013 for the Q1 and Q2 surveys, and for 1996 - 2013 for the Q3 and Q4 surveys. The rows labelled 'SPF_c' show average log scores when the SPF variances are scaled to optimise the log score for the years 1982 - 1996 (Q1 and Q2 surveys) or 1981 - 1995 (Q3 and Q4 surveys). The rows prefixed by 'DM:' record the p-values of Diebold-Mariano tests of equal predictive ability on log score, computed such that values close to 1 reject in favour of the SPF forecasts, and values close to zero reject in favour of the benchmark forecasts.
The table also reports the in-sample estimates of λ which are used to scale the 'out-of-sample' variances.
In all cases the standard deviation of the aggregate SPF histogram is taken to be the average of the individual standard deviations.

9 Not For Publication Appendix

The tables in the appendix provide the detailed results which are summarized in section 6.

Table 9 reports results for the aggregate histograms when the forecast sample begins in 1992:Q1, and tables 10 and 11 report the results for the individual respondents.

Finally, table 12 reports results for the aggregate SPF histograms centred on the median point predictions, and tables 13 and 14 report the same information for the individuals.

Table 9: Aggregate Density Forecasts: SPF & Benchmarks, 1992:Q1-

		Evaluation based on z^*						Comparison (Eqn. 6)
		Ind.	Eqn (2)	$(0,1)$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\sigma}^2$	
Output Growth								
SPF	1	0.16	0.53	0.88	0.06	0.31	0.83	0.99
SPF (no disag.)	1	0.20	0.51	0.71	0.08	0.29	1.11	1.98
BM	1	0.20	0.00	0.00	0.04	0.29	0.22	-
BM (truly uncond.)	1	0.11	0.29	0.55	0.06	0.36	1.21	0.06
SPF	2	0.15	0.07	0.08	0.01	0.34	0.39	1.00
SPF (no disag.)	2	0.19	0.08	0.08	0.01	0.31	0.40	1.00
BM	2	0.82	0.00	0.00	-0.01	0.05	0.21	-
BM (truly uncond.)	2	0.13	0.00	0.00	0.05	0.34	2.97	0.04
SPF	3	0.06	0.00	0.00	0.17	0.43	0.23	0.17
SPF (no disag.)	3	0.05	0.00	0.00	0.16	0.45	0.24	0.34
BM	3	0.59	0.07	0.03	0.20	0.13	0.40	-
BM (truly uncond.)	3	0.12	0.00	0.00	0.05	0.35	8.14	0.03
SPF	4	0.30	0.00	0.00	0.16	0.23	0.21	0.00
SPF (no disag.)	4	0.49	0.01	0.01	0.15	0.16	0.28	0.00
BM	4	0.85	0.02	0.01	0.31	-0.05	0.33	-
BM (truly uncond.)	4	0.11	0.00	0.00	0.00	0.36	45.31	0.03
Inflation								
SPF	1	0.14	0.02	0.02	-0.08	0.35	0.32	0.20
SPF (no disag.)	1	0.21	0.14	0.15	-0.11	0.29	0.47	0.56
BM	1	0.15	0.23	0.33	0.07	0.32	0.55	-
BM (truly uncond.)	1	0.00	0.00	0.00	-0.22	0.79	1.27	0.00
SPF	2	0.02	0.00	0.00	-0.14	0.50	0.14	0.00
SPF (no disag.)	2	0.03	0.00	0.00	-0.16	0.48	0.16	0.00
BM	2	0.19	0.04	0.04	-0.07	0.29	0.35	-
BM (truly uncond.)	2	0.00	0.00	0.00	-0.26	0.76	2.12	0.00
SPF	3	0.75	0.00	0.00	-0.34	-0.08	0.11	0.00
SPF (no disag.)	3	0.96	0.00	0.00	-0.35	-0.01	0.12	0.00
BM	3	0.78	0.00	0.00	-0.16	0.07	0.23	-
BM (truly uncond.)	3	0.00	0.00	0.00	-0.50	0.70	5.84	0.00
SPF	4	0.99	0.00	0.00	-0.36	0.00	0.11	0.00
SPF (no disag.)	4	0.96	0.00	0.00	-0.41	0.01	0.15	0.00
BM	4	0.25	0.04	0.03	-0.21	0.26	0.40	-
BM (truly uncond.)	4	0.00	+DEN	+DEN	-1.06	0.69	24.78	0.00

The table records the results of evaluating the densities using tests related to Eqn. (2), and comparisons based on Diebold-Mariano tests of equal predictive ability using the KLIC loss function, Eqn. (6).
The first column denotes the survey quarter, whereby '1' indicates a first quarter of the year survey, and a forecast horizon of 4 quarters, and '4' a fourth quarter survey (and a horizon of 1 quarter). The column headed 'Ind.' is the p-value of a test for independence: in terms of Eqn. (2) the test is based on $-2(L\left(\tilde{\mu},\tilde{\sigma}^2,0\right)-L\left(\hat{\mu},\hat{\sigma}^2,\hat{\rho}\right))$, where $\tilde{\mu},\tilde{\sigma}^2$ denote MLEs with $\rho=0$ imposed.
The next column is the three-degree of freedom test in Eqn. (2), and the column headed $(0,1)$ tests for zero-mean and unit-variance with a maintained hypothesis of independence. The next 3 columns report the estimates of the unrestricted AR(1).
The final column is the p-value of the test of the SPF against the Benchmark densities (that condition on location) using Eqn. (6). We also compare the two sets of Benchmark densities one against the other.

Table 10: Evaluation of Individual Respondents' Output Histograms, 1992:Q1-

			Evaluation based on z^*			Comparison (Eqn. 6)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	
421	1	19	0.57	0.14	1.31	0.78
431	1	17	0.20	0.36	1.38	0.39
446	1	18	0.16	0.21	1.68	0.61
426	1	20	0.01	0.49	1.73	0.36
411	1	16	0.01	0.36	2.28	0.23
484	1	17	0.00	0.79	0.89	0.12
420	1	16	0.00	0.01	3.11	0.23
433	1	16	0.00	0.62	2.62	0.16
421	2	20	1.00	-0.01	1.00	0.31
446	2	20	0.81	0.13	0.92	0.62
431	2	17	0.53	0.14	0.70	0.89
426	2	18	0.25	0.21	0.60	0.81
463	2	17	0.22	-0.24	0.58	0.58
484	2	16	0.18	0.29	0.57	0.09
433	2	17	0.14	0.45	1.26	0.61
421	3	16	0.39	0.34	1.03	0.06
420	3	20	0.23	0.16	0.58	0.36
407	3	17	0.10	0.48	1.32	0.85
433	3	18	0.06	0.48	0.67	0.96
426	3	20	0.06	0.28	0.50	0.03
446	3	18	0.02	0.37	0.41	0.02
421	4	19	0.70	0.16	0.86	0.01
411	4	16	0.46	0.24	0.74	0.01
446	4	17	0.43	0.31	0.99	0.00
84	4	17	0.27	0.29	1.41	0.01
433	4	17	0.22	0.42	1.08	0.00
472	4	16	0.06	0.29	0.44	0.00
431	4	16	0.04	0.62	1.14	0.01
407	4	17	0.04	0.62	1.05	0.01
426	4	19	0.02	0.20	0.38	0.00
420	4	18	0.00	0.44	0.24	0.00
463	4	17	0.00	0.09	0.15	0.00

The table records the results of evaluating the densities of individual respondents using tests related to Eqn. (2), and comparisons based on Diebold-Mariano tests of equal predictive ability using the KLIC loss function, Eqn. (6), to compare each individual's forecasts against the Benchmark (where the Benchmark conditions on location).
'N' is the number of forecast densities by the individual 'id' made in response to 'Qtr' surveys. The column headed $(0,1)$ tests for zero-mean and unit-variance with a maintained hypothesis of independence, and the next 2 columns report the estimates of mean and variance.
The final column are the p-values of the test of an SPF individual against the Benchmark densities using Eqn. (6). We consider all respondents who made more than 15 forecasts of a given horizon.

Table 11: Evaluation of Individual Respondents' Inflation Histograms, 1992-

			Evaluation based on z^*			Comparison (Eqn. 6)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	
411	1	17	0.52	0.23	0.80	0.22
446	1	18	0.23	-0.16	0.56	0.14
426	1	20	0.15	-0.27	1.57	0.13
420	1	16	0.12	-0.20	1.83	0.11
433	1	17	0.12	0.13	1.85	0.14
484	1	17	0.11	-0.19	0.47	0.01
421	1	17	0.02	-0.36	0.41	0.03
431	1	18	0.00	-0.10	0.22	0.55
431	2	16	0.37	-0.35	1.00	0.05
433	2	17	0.25	-0.32	0.69	0.07
446	2	20	0.06	-0.33	0.52	0.00
426	2	18	0.02	-0.63	0.74	0.00
463	2	18	0.02	-0.38	0.41	0.00
421	2	18	0.00	-1.11	1.79	0.04
426	3	20	0.01	-0.65	0.68	0.00
420	3	20	0.00	-0.44	0.35	0.00
407	3	16	0.00	-0.86	0.60	0.00
446	3	18	0.00	-0.36	0.24	0.00
433	3	18	0.00	-0.30	0.22	0.00
84	4	17	0.31	-0.37	1.07	0.00
411	4	16	0.30	-0.13	0.57	0.00
431	4	16	0.23	-0.42	1.09	0.00
463	4	17	0.02	-0.28	0.36	0.00
426	4	18	0.01	-0.44	0.40	0.00
433	4	16	0.01	-0.69	1.58	0.00
446	4	18	0.00	-0.47	0.34	0.00
421	4	18	0.00	-0.81	0.81	0.00
407	4	16	0.00	-0.75	1.77	0.00
420	4	18	0.00	-0.51	0.23	0.00

The table records the results of evaluating the densities of individual respondents using tests related to Eqn. (2), and comparisons based on Diebold-Mariano tests of equal predictive ability using the KLIC loss function, Eqn. (6), to compare each individual's forecasts against the Benchmark (where the Benchmark conditions on location).
'N' is the number of forecast densities by the individual 'id' made in response to 'Qtr' surveys. The column headed $(0,1)$ tests for zero-mean and unit-variance with a maintained hypothesis of independence, and the next 2 columns report the estimates of mean and variance.
The final column are the p-values of the test of an SPF individual against the Benchmark densities using Eqn. (6). We consider all respondents who made more than 15 forecasts of a given horizon.

Table 12: Aggregate Density Forecasts: SPF & Benchmarks, Centred on Median Point Predictions

		Evaluation based on z^*						Comparison (Eqn. 6)
		Ind.	Eqn (2)	$(0,1)$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\sigma}^2$	
Output Growth								
SPF	1	0.24	0.27	0.29	0.23	0.24	0.72	1.00
SPF	2	0.71	0.01	0.01	0.01	0.07	0.37	1.00
SPF	3	0.35	0.00	0.00	0.10	0.18	0.31	0.00
SPF	4	0.51	0.00	0.00	0.08	0.12	0.08	0.00
Inflation								
SPF	1	0.03	0.00	0.00	0.03	0.40	0.25	0.66
SPF	2	0.13	0.00	0.00	-0.09	0.27	0.11	0.00
SPF	3	0.23	0.00	0.00	-0.13	-0.20	0.05	0.00
SPF	4	0.15	0.00	0.00	-0.06	0.26	0.02	0.00

The table records the results of evaluating the densities using tests related to Eqn. (2), and comparisons based on Diebold-Mariano tests of equal predictive ability using the KLIC loss function, Eqn. (6).
The first column denotes the survey quarter, whereby '1' indicates a first quarter of the year survey, and a forecast horizon of 4 quarters, and '4' a fourth quarter survey (and a horizon of 1 quarter). The column headed 'Ind.' is the p-value of a test for independence: in terms of Eqn. (2) the test is based on $-2(L\left(\tilde{\mu},\tilde{\sigma}^2,0\right)-L\left(\hat{\mu},\hat{\sigma}^2,\hat{\rho}\right))$, where $\tilde{\mu},\tilde{\sigma}^2$ denote MLEs with $\rho=0$ imposed.
The next column is the three-degree of freedom test in Eqn. (2), and the column headed $(0,1)$ tests for zero-mean and unit-variance with a maintained hypothesis of independence. The next 3 columns report the estimates of the unrestricted AR(1).
The final column is the p-value of the test of the SPF against the Benchmark densities (that condition on location) using Eqn. (6).

Table 13: Evaluation of Individual Respondents' Output Growth Histograms, Centred on Point Predictions

			Evaluation based on z^*			Comparison (Eqn. 6)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	
421	1	20	0.65	-0.06	1.31	0.70
431	1	18	0.45	0.21	0.73	0.99
446	1	18	0.36	0.17	1.47	0.84
411	1	17	0.33	0.14	1.55	0.65
420	1	16	0.23	-0.06	1.73	0.82
484	1	17	0.07	0.55	1.14	0.11
99	1	16	0.06	-0.30	1.91	0.66
428	1	16	0.04	0.64	1.14	0.77
426	1	21	0.00	0.56	1.80	0.31
433	1	17	0.00	0.53	2.40	0.20
407	1	16	0.00	0.65	2.76	0.29
20	1	17	0.00	0.00	4.10	0.02
446	2	20	0.41	0.07	0.64	0.95
431	2	18	0.40	0.07	0.62	0.92
421	2	21	0.23	-0.05	0.56	0.88
463	2	17	0.16	-0.20	0.52	0.78
411	2	16	0.13	-0.19	0.48	0.89
433	2	18	0.10	0.50	1.00	0.74
99	2	16	0.10	-0.13	0.43	1.00
407	2	16	0.05	0.02	2.14	0.62
426	2	19	0.02	0.27	0.40	0.94
484	2	16	0.01	0.12	0.29	0.55
20	2	23	0.00	-0.25	2.22	0.12
65	2	18	0.00	0.32	2.45	0.16
20	3	21	0.36	-0.31	1.07	0.13
421	3	17	0.28	0.15	0.58	0.38
407	3	19	0.23	0.34	1.30	0.75
433	3	19	0.13	0.46	0.87	0.84
420	3	20	0.06	0.25	0.48	0.51
84	3	19	0.05	0.22	1.91	0.13
446	3	18	0.01	0.29	0.34	0.03
65	3	17	0.01	0.37	2.26	0.03
426	3	21	0.00	0.38	0.33	0.03
84	4	27	0.73	0.14	1.09	0.00
433	4	18	0.38	0.33	0.97	0.00
431	4	17	0.34	0.34	0.86	0.03
446	4	17	0.33	0.27	0.70	0.00
20	4	20	0.22	-0.30	1.40	0.00
407	4	19	0.17	0.42	0.88	0.02
99	4	18	0.06	0.49	1.44	0.01
472	4	16	0.04	0.22	0.38	0.00
421	4	21	0.02	0.23	0.41	0.01
463	4	17	0.00	0.09	0.11	0.00
411	4	18	0.00	0.13	0.10	0.00
420	4	20	0.00	0.19	0.10	0.00
426	4	20	0.00	0.15	0.10	0.00

The table records the results of evaluating the densities of individual respondents using tests related to Eqn. (2), and comparisons based on Diebold-Mariano tests of equal predictive ability using the KLIC loss function, Eqn. (6), to compare each individual's forecasts against the Benchmark (where the Benchmark conditions on location).
'N' is the number of forecast densities by the individual 'id' made in response to 'Qtr' surveys. The column headed $(0,1)$ tests for zero-mean and unit-variance with a maintained hypothesis of independence, and the next 2 columns report the estimates of mean and variance.
The final column are the p-values of the test of an SPF individual against the Benchmark densities using Eqn. (6). We consider all respondents who made more than 15 forecasts of a given horizon.

Table 14: Evaluation of Individual Respondents' Inflation Histograms, Centred on Point Predictions

			Evaluation based on z^*			Comparison (Eqn. 6)
id.	Qtr.	N.	$(0,1)$	$\hat{\mu}$	$\hat{\sigma}^2$	
420	1	16	0.55	-0.02	1.44	0.18
426	1	21	0.31	-0.04	0.60	0.34
99	1	16	0.20	-0.04	0.49	0.89
484	1	17	0.06	0.12	0.40	0.03
433	1	18	0.06	0.09	1.99	0.10
411	1	18	0.03	0.49	0.55	0.31
446	1	18	0.02	0.19	0.36	0.33
431	1	19	0.01	0.24	0.30	0.26
421	1	18	0.00	-0.11	0.26	0.05
407	1	16	0.00	-0.73	2.51	0.04
20	1	16	0.00	0.19	7.36	0.14
99	2	17	0.96	-0.06	0.94	0.07
411	2	16	0.68	-0.19	1.16	0.01
433	2	18	0.12	-0.30	0.56	0.10
463	2	18	0.10	-0.31	0.54	0.00
65	2	17	0.09	-0.47	1.41	0.12
407	2	16	0.03	-0.65	1.06	0.03
426	2	19	0.02	-0.34	0.40	0.00
446	2	20	0.00	-0.05	0.25	0.00
431	2	17	0.00	-0.04	0.13	0.00
421	2	19	0.00	-0.27	0.15	0.00
20	2	23	0.00	-0.85	2.18	0.00
84	3	20	0.10	-0.38	0.64	0.00
20	3	21	0.00	-0.70	1.33	0.00
65	3	17	0.00	-0.36	2.49	0.01
407	3	18	0.00	-0.71	0.35	0.00
426	3	21	0.00	-0.18	0.15	0.00
433	3	19	0.00	-0.32	0.11	0.00
420	3	20	0.00	-0.09	0.08	0.00
446	3	18	0.00	-0.03	0.05	0.00
431	4	17	0.41	-0.32	1.01	0.00
84	4	28	0.34	-0.19	1.32	0.00
99	4	18	0.07	-0.54	1.18	0.00
421	4	20	0.06	-0.46	0.69	0.00
463	4	17	0.06	-0.54	1.27	0.00
20	4	19	0.02	-0.65	0.94	0.00
411	4	18	0.01	-0.16	0.31	0.00
433	4	17	0.01	-0.72	1.39	0.00
407	4	18	0.00	-0.68	1.62	0.00
446	4	18	0.00	-0.41	0.31	0.00
426	4	19	0.00	-0.09	0.14	0.00
420	4	20	0.00	-0.18	0.10	0.00

The table records the results of evaluating the densities of individual respondents using tests related to Eqn. (2), and comparisons based on Diebold-Mariano tests of equal predictive ability using the KLIC loss function, Eqn. (6), to compare each individual's forecasts against the Benchmark (where the Benchmark conditions on location).
'N' is the number of forecast densities by the individual 'id' made in response to 'Qtr' surveys. The column headed $(0,1)$ tests for zero-mean and unit-variance with a maintained hypothesis of independence, and the next 2 columns report the estimates of mean and variance.
The final column are the p-values of the test of an SPF individual against the Benchmark densities using Eqn. (6). We consider all respondents who made more than 15 forecasts of a given horizon.