Published in final edited form as:

Int J Forecast. 2015 ; 31(1): 113–129. doi:10.1016/j.ijforecast.2014.03.005.

Testing the Value of Probability Forecasts for Calibrated Combining

Kajal Lahiri, Huaming Peng, and Yongchen Zhao

Department of Economics, University at Albany, SUNY, Albany, New York, 12222, USA

Abstract

We combine the probability forecasts of a real GDP decline from the U.S. Survey of Professional Forecasters, after trimming the forecasts that do not have “value”, as measured by the Kuiper Skill Score and in the sense of Merton (1981). For this purpose, we use a simple test to evaluate the probability forecasts. The proposed test does not require the probabilities to be converted to binary forecasts before testing, and it accommodates serial correlation and skewness in the forecasts. We find that the number of forecasters making valuable forecasts decreases sharply as the horizon increases. The beta-transformed linear pool combination scheme, based on the valuable individual forecasts, is shown to outperform the simple average for all horizons on a number of performance measures, including calibration and sharpness. The test helps to identify the good forecasters ex ante, and therefore contributes to the accuracy of the combined forecasts.

1. Introduction

Currently, more than forty-five years of quarterly expert forecasts for a number of U.S. macroeconomic variables are available in the U.S. Survey of Professional Forecasters (SPF).

¹ In particular, the SPF probabilities of real GDP declines for the current quarter and the next four quarters have attracted special interest from economists and the mass media alike. The New York Times columnist David Leonhardt (September 1, 2002) called the one-quarter-ahead GDP decline probability the “Anxious Index”. Drawing on methodologies developed in statistics, the atmospheric sciences, and psychology, economists have studied the quality and characteristics of these subjective probabilities, and have reached a certain broad consensus.² First, a simple average of the individual probabilities seems to encompass all of the information embedded in the individual forecasts (Clements & Harvey 2011); and second, these average forecast probabilities do not seem to have any predictive power

© 2014 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

klahiri@albany.edu (Kajal Lahiri), hpeng@albany.edu (Huaming Peng), hzhao@albany.edu (Yongchen Zhao)

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

¹The European Central Bank, the Banco de Mexico and the Reserve Bank of India have also been collecting similar forecasts under the same name. Thus, any methodological improvement in the use of the U.S. data will be useful for a number of countries as well.

²See for instance Braun & Yaniv (1992), Graham (1996), Lahiri & Wang (2006), Clements (2008), Clements (2009), Engelberg et al. (2010), Clements (2011), Clements & Harvey (2011), Galbraith & van Norden (2011), Galbraith & van Norden (2012), and Lahiri & Wang (2013).

beyond the second quarter (Lahiri & Wang 2013). Many researchers have noted the limitations of these average probabilities in forecasting economic downturns. Stock & Watson (2003) point out that the SPF could not foresee the 2001 recession: the signal came with a lag in 2001:Q4, when the negative growth period had already passed. Rudebusch & Williams (2009) show that the SPF participants do not seem to use the information in the yield spread (i.e., the difference between long-term and short-term interest rates) in forecasting recessions, despite it having been well-known since at least the 1980s that the yield spread is useful in forecasting real GDP growth and recessions. However, Lahiri et al. (2013) find that, when averaged over the relatively better forecasters, the combined SPF forecasts do incorporate the information from the yield spread, as well as that from a myriad of other economic indicators. Galbraith & van Norden (2012) use the pioneering “fan charts” of the Bank of England to calculate the probability that the annual rates of inflation and output growth will exceed given thresholds. They find a serious departure of these aggregate forecasts from perfect calibration and reasonable sharpness.

The main motivation of this paper comes from a recent finding by Ranjan & Gneiting (2010) that the linear opinion pool is sub-optimal in terms of calibration and sharpness. Following their suggestion, we use the non-linearly recalibrated beta-transformed linear pool (BLP) to determine whether the performances of aggregated SPF forecasts can be improved. As the fallout from the recent recession of 2007–2009 has painfully reminded us again, even a small improvement in our capability to foresee recessions is of enormous benefit to any modern society. Our second motivation is derived from the novel methodological approach to forecast combination by Aiolfi & Timmermann (2006). Following their approach, and due to the huge amount of missing data in the SPF, we first sort all of the forecasters into four clusters based on their past performances. We pool the forecasts within each cluster, and follow this with an application of the BLP methodology to the cluster-level aggregates, in order to obtain an improved combined forecast. However, before constructing the clusters, we trim the forecasters whose forecasts are found to have no value in the sense of the Kuiper Skill Score (KSS), which is a sample analog of the concept of “value” in Merton (1981). For assessing the “value” of the forecasts, we propose to use a Welch-type test³ that explicitly addresses the potential issue of serial correlation and skewness.

The plan of the paper is as follows. In Section 2, we describe the nature and characteristics of the SPF probability forecasts, which naturally leads on to methodological approach undertaken in this paper. In Section 3, we describe the test for zero forecast value and discuss its robustness to serial correlation and skewness. Section 4 reports our empirical results, including the test results, the beta-transformed linear pool methodology, the combination of the valuable forecasts, and a detailed evaluation of the combined forecasts in- and out-of-sample. Section 5 concludes.

³The Welch-type t test is a modified Student's t -test which is designed to test the equality of two means with possibly unequal

variances. That is, $t_w = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$, where $\hat{\mu}_i$, s_i^2 and N_i are the sample mean, sample variance and sample size respectively for $i = 1, 2$, and the degree of freedom ν is given approximately by $\nu \approx \left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} \right)^2 / \left(\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)} \right)$.

2. Probability forecasts in SPF

We use the probability forecasts of real GDP declines from the U.S. Survey of Professional Forecasters (SPF). The SPF is the oldest quarterly survey of macroeconomic forecasts in the United States.⁴ Since 1990Q2, the survey has been being administered by the Federal Reserve Bank of Philadelphia. Survey respondents are asked to supply point and density forecasts for a wide range of variables covering output, prices, and employment situations. The survey is used widely by researchers. Examples of recent work using data from the survey include the studies by Capistran & Timmermann (2009), Clements (2011) and Lahiri & Wang (2013). We examine the probability forecasts of real GDP declines for the current quarter (denoted $h = 0$)⁵ and the following four quarters ($h = 1, \dots, 4$). The equally-weighted average of individual forecasts, i.e., the equally-weighted linear opinion pool (ELP), is used as a benchmark when evaluating the performances of the combined forecasts we construct. Our data spans the 174 quarters from 1968Q4 to 2012Q1. A total of 32,379 forecasts available from 426 forecasters for 5 horizons.

Frequent data revisions and definitional changes affect real GDP values after their initial release. In practice, such changes are rarely predictable they are announced. The choice of actual values between an earlier data vintage and the latest vintage depends crucially on the objective function of a forecasting client. We construct the binary outcome series (the actual values) y_t using the first vintage in the Real-Time Data Set for Macroeconomists available from the Federal Reserve Bank of Philadelphia.⁶ For a discussion of the real time data issues involved in evaluating SPF forecasts, see Stark (2010). The same approach is taken by, among others, Clements & Harvey (2010). Our binary outcome series contains 24 quarters of real GDP declines, which is about 14% of the sample. Since a real GDP decline is a relatively uncommon event, special considerations are needed in forecast evaluation. Examples of recent work focusing on evaluating probability forecasts include the studies Lahiri & Wang (2013) and Galbraith & van Norden (2012)

The SPF contains a large amount of missing data, see Capistran & Timmermann (2009) and Lahiri et al. (2012). In our case, more than 91% of the data is missing.⁷ To make the situation more complicated, after the Federal Reserve Bank took over the survey, most of the old forecasters stopped forecasting and many new forecasters joined the sample. As a result, for most of the old forecasters, no forecast is available after 1990Q2, before which there is nothing available from the newly joined forecasters. To address this issue, we split the sample at 1990Q2, thus creating two subsamples.⁸ Subsample 1 (the earlier subsample) has 275 forecasters. Subsample 2 has 158 forecasters. To further limit the amount of missing data, we impose a participation requirement. A simple requirement asking for more than a certain number of forecasts from a forecaster would not work in our case. This is because,

⁴The survey was introduced in 1968 by the American Statistical Association and the National Bureau of Economic Research. For more information and background about the survey, see Croushore (1993). The survey itself can be accessed from <http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/>.

⁵The "current quarter" refers to the quarter in which the survey is conducted.

⁶A robustness check was conducted using the latest vintage, and our main conclusions re-mained the same. Results from the robustness check are available from the authors.

⁷A fully balanced panel with all forecasters who have ever participated and all quarters would have 426 forecasters $\times$ 174 quarters $\times$ 5 horizons = 370, 620 forecasts; but we have only 32, 379 forecasts.

⁸Using the sample in its entirety does not alter our results qualitatively.

for many forecasters who satisfy such a requirement, all of their forecasts come from quarters without a real GDP decline. Therefore, we impose the requirement separately on the quarters with and without a real GDP decline: for each forecaster in subsample 1 (2), we require at least 7 (3) forecasts from quarters with a decline in real GDP and at least 7 (3) forecasts from quarters with a growth in real GDP. This leaves us around 50 forecasters for each horizon.⁹ Imposing this requirement decreases the amount of missing data from 86% to 56% for subsample 1 and from 78% to 55% for subsample 2.

Table 1 shows several descriptive statistics on individual QPSs¹⁰ and skill scores for those satisfying the participation requirement. As the horizon increases, the mean value of the individual QPS deteriorates from 0.12 for current-quarter forecasts to 0.25 for four-quarter-ahead forecasts in subsample 1, and from 0.07 to 0.13 in subsample 2. The standard deviation of the QPS scores increases slightly as well: from 0.04 to around 0.06 as forecast horizon increases in both subsamples. In general, the forecasting performances of these frequent forecasters seem to be better in the later years (subsample 2) than in the earlier years. Across horizons, for both subsamples, a clear deterioration in performance is observed at around the two-quarter horizon.

The Kuiper Skill Scores (KSS)¹¹ for forecasters satisfying the participation requirements are shown in Figure 1a for subsample 1 and in Figure 1b for sub-sample 2. These individual Kuiper scores are ranked and plotted on the vertical axis for all five horizons, with the number of forecasters being plotted on the horizontal axis. Each point in the graph represents one forecaster. Points joined by a curve represent forecasters who are forecasting for the same horizon. For each horizon, forecasters are ranked based on their KSS. The longer a curve, i.e., the more points connected by a curve, the more forecasters there are for the horizon represented by the curve. Ideally, the curves for shorter horizons should sit above the curves for longer horizons, representing a deterioration of KSS as horizon increases. The steeper a curve is, the more rapidly KSS deteriorates from the best forecaster to the worst forecaster for a given horizon. We see from the top plot in both Figures 1a and 1b that the forecasts for the current quarter are significantly better than those for the remaining quarters. The performances of these forecasts deteriorate greatly from the current quarter to one quarter ahead for both subsamples, but only modestly from one to two quarters ahead for subsample 1. Little deterioration in KSS is observed between the three- and four-quarter-ahead forecasts. The best forecasters for the two-, three-, and four-quarter-ahead horizons have almost the same performances. A zero or negative KSS is first observed at the two-quarter horizon. Within each quarter, the KSSs of the best few forecasters and

⁹We also carried out our analysis with this requirement increased to 10 (5) and decreased to 5 (2) for subsample 1 (2), and reached similar conclusions. Note that a requirement higher than 7 would filter out all forecasters joining the survey after the 1990s, since there have not been more than 7 quarters with a real GDP decline in the later period.

¹⁰The Brier's Quadratic Probability Score (QPS) is a commonly used measure of forecast accuracy, being a probability analog of the mean squared error. See Lahiri & Wang (2013) for its definition, decomposition, and application in evaluating probability forecasts.

¹¹The Kuiper (or Peirce) skill score is defined as $KSS(\omega_t) = \frac{\sum_{t=1}^T \mathbb{I}_{\{\hat{p}_t > \omega_t, y_t=1\}}}{\mathbb{I}_{\{y_t=1\}}} - \frac{\sum_{t=1}^T \mathbb{I}_{\{\hat{p}_t > \omega_t, y_t=0\}}}{\mathbb{I}_{\{y_t=0\}}}$ for some threshold probability value $\omega_t > 0$. Assume also that $\mathbb{I}_{\{y_t=1\}}$, $\mathbb{I}_{\{y_t=0\}}$, $\mathbb{I}_{\{\hat{p}_t > \omega_t, y_t=1\}}$, and $\mathbb{I}_{\{\hat{p}_t > \omega_t, y_t=0\}}$ are stationary and ergodic over time; then, it can be shown that the probability limit of KS converges to $P(\hat{p}_t > \omega_t | y_t=1) - P(\hat{p}_t > \omega_t | y_t=0)$, which is a function of ω_t .

those of the worst few forecasters are widely different, while the KSSs for forecasters in the middle are rather similar.

To further illustrate some common characteristics of these probability forecasts, we collect all of the individual forecasts horizon by horizon, conditioned on the outcome, and fit a beta distribution to them. We present these fitted densities in Figure 2, where two such conditional densities are plotted for each horizon, together with the associated mean values (shown by vertical lines, with the mean value for the periods with a real GDP decline on the right). These figures show a diminishing wedge between the two densities as the horizon increases. For the four-quarter-ahead forecasts (horizon 4), there is virtually no difference between the two conditional densities. Another prominent feature of these probability forecasts is that they are highly skewed. From horizons 0 to 4, for periods with a real GDP decline, the skewness values are -0.31, 0.23, 0.56, 0.81, and 0.91. From horizons 0 to 4, for periods of real GDP growth, the values are 1.06, 1.01, 0.95, and 0.93. This observation underscores the need for to skewness in the data when testing the value of the forecasts.

3. A test of the probability forecasts

Consider the binary variable of interest y_t , where

$$y_t = \begin{cases} 1, & \text{if an event } Y \text{ occurs at time } t \\ 0, & \text{otherwise.} \end{cases}$$

Let a forecast be made at time $t - h$ of the binary variable realization at time t , where h is the forecast horizon. Let $\hat{y}_t \in \{0, 1\}$ be a binary forecast of y_t , and $\hat{p}_t$ be a forecast of the probability of $\hat{y}_t$, since it gives the distribution of all possible outcomes. Hence, one can easily construct $\hat{y}_t$ using $\hat{p}_t$ while in general, one cannot do the reverse. A simple and popular way to construct $\hat{y}_t$ is to use the indication function, that is, to set $\hat{y}_t = 1_{\{\hat{p}_t > \omega_t\}}$ for some threshold value $\omega_t \in (0, 1)$, which makes $\hat{y}_t$ a function of the nuisance parameter ω_t .

Given $\hat{p}_t$ and the realized value y_t , it is interesting to determine whether the forecast probability $\hat{p}_t$ has any economic significance. According to Merton (1981), if $\hat{y}_t = 1_{\{\hat{p}_t > \omega_t\}}$, then the forecast probability $\hat{p}_t$ has no economic value whenever

$$P(\hat{p}_t > \omega_t | y_t = 1) + P(\hat{p}_t \leq \omega_t | y_t = 0) = 1, \quad (1)$$

and $\hat{p}_t$ is valuable if and only if

$$P(\hat{p}_t > \omega_t | y_t = 1) + P(\hat{p}_t \leq \omega_t | y_t = 0) > 1. \quad (2)$$

Merton (1981) argues that the larger the value of the sum $P(\hat{p}_t > \omega_t | y_t = 1) + P(\hat{p}_t \leq \omega_t | y_t = 0)$, the more valuable the forecasts $\hat{p}_t$.¹² From the property of discrete conditional probability, it can be seen that the two conditions in Eqs. (1) and (2) are equivalent to

¹²See Stekler(1994) and Schnader & Stekler (1990) for an early explication of the concept and an application to macroeconomic forecasts.

$$P(\hat{p}_t > \omega_t | y_t = 1) - P(\hat{p}_t > \omega_t | y_t = 0) = 0, \quad (3)$$

and

$$P(\hat{p}_t > \omega_t | y_t = 1) - P(\hat{p}_t > \omega_t | y_t = 0) > 0, \quad (4)$$

respectively. Note that $P(\hat{p}_t > \omega_t | y_t = 1)$ and $P(\hat{p}_t > \omega_t | y_t = 0)$ are the hit rate and the false alarm rate (or their respective probability limits), because $\hat{y}_t = 1$ whenever $\hat{p}_t > \omega_t$. Thus, the forecast probability $\hat{p}_t$ has positive economic value if and only if the (limit of the) Kuiper score, defined as the difference between the hit rate and the false alarm rate, is positive. The application of the Kuiper as an indication of the economic value of forecasts of binary events has also been justified by Granger & Pesaran (2000). Intuitively, in order for a to be valuable in Merton's sense, it is both necessary and sufficient that it should outperform some random forecasts with an average of a zero Kuiper score.

However, the (limit of the) Kuiper score, as on the left hand side of Eqs. (3) and (4), depends on the threshold probability ω_t . The Kuiper score calculated using the threshold ω_t clearly has a “pointwise” property. Since the test statistic, and statistical inference in general, vary with ω_t , it would only determine the economic value of forecasts conditional on the specific ω_t . To construct a hypothesis test of economic value that is globally valid for the whole range of ω_t , we notice that

$$\int_0^1 [P(\hat{p}_t > \omega_t | y_t = 1) - P(\hat{p}_t > \omega_t | y_t = 0)] d\omega_t = E(\hat{p}_t | y_t = 1) - E(\hat{p}_t | y_t = 0).$$

As a result, testing the economic value of $\hat{p}_t$ globally boils down to testing the null hypothesis

$$H_0: E(\hat{p}_t | y_t = 1) - E(\hat{p}_t | y_t = 0) = 0, \quad (5)$$

against the alternative hypothesis

$$H_1: E(\hat{p}_t | y_t = 1) - E(\hat{p}_t | y_t = 0) > 0. \quad (6)$$

One objective of this paper is to construct a simple test for the economic value of a set of probability forecasts without converting the probabilities p_t to binary forecasts. For notational simplicity, $p_{1,t}$ and $p_{0,t}$ denote the observed forecast probability conditioned on $y_t = 1$ and $y_t = 0$, respectively, and assume:

Assumption A:

(i) $i = 0, 1$ $p_{i,t}$ is a stationary and α -mixing process, with $\alpha_i(t)$ being of size $(2 + 2\delta)/\delta$, $E p_{i,t} = p_i$ and $E |p_{i,t}|^{8+\delta} < \infty$ for some $\delta > 0$. In addition,

$$|E(\hat{p}_{i,t}^* \hat{p}_{i,t-s}^* \hat{p}_{i,t-\tau}^*)| \leq [\alpha_i(s) \alpha_i(\tau)]^{\frac{\delta}{2+2\delta}} M \text{ for } 0 < s < \tau \text{ and some } 0 < M < \infty, \text{ where } \hat{p}_{i,t}^* = \hat{p}_{i,t} - p_i.$$

(ii) $p_{0,t}$ is independent of $p_{1,s}$ for all (t, s) .

Assumption B: $\omega(j, T_{ri}) = 1 - \frac{j}{T_{ri}+1}, j = 1, 2, \dots, T_{ri}$ for $i = 0, 1$, where T_{ri} is a function of T_i such that $\lim_{T_i \rightarrow \infty} T_{ri}(T_i) = \infty$ and $\lim_{T_i \rightarrow \infty} \frac{T_{ri}}{T_i} = 0$.

Assumption A allows for some weak dependence over time in the probability forecast $\hat{p}_{i,t}$. The size of strong mixing, which controls the rate of dependence, along with the moment condition restriction and the independence between the two processes, is important for deriving the asymptotic normal distribution via the central limit theorem for dependent data.

Assumption B, together with the rate restriction on $E(\hat{p}_{i,t}^* \hat{p}_{i,t-s}^* \hat{p}_{i,t-\tau}^*)$, enables us to obtain HAC estimators for both the long-run variance and skewness in the spirit Newey & West (1987) and Newey & West (1994), and is fairly standard in the literature dealing with HAC estimators. Taken together, Assumptions A and B specify the dynamic properties of the two probability forecast processes and the conditions that allow us to carry out a simple one-sided test for the economic value of $\hat{p}_t$.

A number of statistical tests for the economic value of directional forecasts have been proposed in the literature. Henriksson & Merton (1981) introduced a statistical test of forecasting skill that is closely related to Fisher's for testing the null of independence between two binary variables, cf. Stekler (1994). Under the assumption of serial independence, Pesaran & Timmermann (1992) proposed an asymptotic test for evaluating the economic value of binary forecasts. The asymptotic test is essentially a standardized version of the Kuiper score, as was shown by Granger & Pesaran (2000). Pesaran & Timmermann (1994) generalized the asymptotic test to multi-category variables. Recently, Pesaran & Timmermann (2009) (referred to as PT below) extended the asymptotic test for the economic value of directional forecasts to the more realistic situation of serial correlation. However, as was reported by Pesaran & Timmermann (2009), this test suffers from small sample size distortions. All of these tests the binary point forecasts $\hat{y}_t$. A threshold value ω_t would be required in order to transform $\hat{p}_t$ to $\hat{y}_t$ if the latter is not observed directly. As was pointed out by Stekler (1994), these binary values would ignore the quantitative significance of the probability forecasts. One obvious choice ω_t is 0.5, but other values could also be used. In particular, the threshold value ω_t is often chosen in such a way that some skill score is maximized (Sohn & Park 2008). However, as the optimal threshold ω_t depends on the chosen skill score, it would also depend on the corresponding economic value function or objective function. Moreover, the test statistic and statistical inference may be misleading if the optimal ω_t selected for optimizing the skill score is far from the unknown true value (if it exists). In addition, as ω_t may vary over time, one may not be able to obtain an optimal threshold value for all t . Furthermore, the choice of an optimal threshold ω_t becomes even more complicated if there are many forecasters who have potentially different thresholds. In general, one may need to allow ω_t to vary across forecasters, horizons, and time in order to obtain valid test results using derived binary forecasts instead of the observed probability forecasts.

Our simple z test for Eq. (5) against Eq. (6) avoids the tricky issue of choosing the threshold values. With this aim, we define, for $i = 0, 1$, $\bar{p}_i = \frac{1}{T_1} \sum_{t=1}^{T_1} \hat{p}_{it}$, $\hat{p}_{i,t}^* = \hat{p}_{i,t} - p_i$, $\tilde{p}_{i,t} = \hat{p}_{i,t} - \bar{p}_i$, $s_i^2 = \frac{1}{T_i} \sum_{t=1}^{T_i} \tilde{p}_{i,t}^2 + \frac{1}{T_i} \sum_{s=1}^{T_{ri}} \sum_{t=s+1}^{T_i} \omega(s, T_{ri}) \tilde{p}_{i,t} \tilde{p}_{i,t-s}$ where $T_{ri} = \left[4(T_i/100)^{2/15} \right]$, and $\sigma_i^2 = \lim_{T_i \rightarrow \infty} \text{Var}(\bar{p}_i)$. Then, under the null hypothesis that $p_1 = p_0$, the test statistic is

$$z_{T_1, T_0} = \frac{\bar{P}_1 - \bar{P}_0}{s} \rightarrow_d N(0, 1) \quad \text{as } T_1 \rightarrow \infty, T_0 \rightarrow \infty,$$

where $s = \sqrt{\frac{s_0^2}{T_0-1} + \frac{s_1^2}{T_1-1}}$. The proof of the asymptotic limit distribution follows from

Corollary 5.3 (ii) of Hall & Heyde (1980) and the fact that $s / \sqrt{\frac{s_0^2}{T_0-1} + \frac{s_1^2}{T_1-1}} \rightarrow_p 1$ as $T_1 \rightarrow \infty$, $T_0 \rightarrow \infty$, which is implied by the consistency of s_0^2 and s_1^2 , established in Theorem 2 of Newey & West (1987).

It seems appealing to use z_{T_1, T_0} directly to test whether p_t has economic value. However, it has been well documented that this type of test may be biased and not robust when the underlying true distribution of p_{it} is skewed like the SPF data we use here.¹³ A better Type-I error coverage could be achieved by utilizing the transformation method of Johnson (1978) or Hall (1992) to remove the biasing effect of skewness (see for example Luh & Guo 1999, Guo & Luh 2000). To this end, we follow Newey & West (1987) and Newey & West (1994) in computing the heteroscedasticity and autocorrelation consistent skewness. That the skewness of $T_i^{1/3} \bar{p}_i$ is estimated by

$$\begin{aligned} \hat{S}_{k, T_{ri}} &= \frac{1}{T_i} \sum_{t=1}^{T_i} \tilde{p}_{i,t}^3 \\ &+ \frac{3}{T_i} \sum_{s=1}^{T_{ri}} \sum_{t=s+1}^{T_i} \omega(s, T_{ri}) \tilde{p}_{i,t}^2 \tilde{p}_{i,t-s} \\ &+ \frac{3}{T_i} \sum_{s=1}^{T_{ri}-1} \sum_{t=s+1}^{T_i} \omega(s, T_{ri}) \tilde{p}_{i,t}^2 \tilde{p}_{i,t-s} \\ &+ \frac{6}{T_i} \sum_{\tau=1}^{T_{ri}} \sum_{s=\tau+1}^{T_{ri}} \sum_{t=s+1}^{T_i} \omega(\tau, T_{ri}) \omega(s, T_{ri}) \tilde{p}_{i,t-s} \tilde{p}_{i,t-s-\tau}, \end{aligned}$$

in an attempt to correct the effects of serial correlation in the probability forecasts. Compared to the weight used in the HAC variance estimator, the weight in the last term of the HAC skewness estimator consists of an addition penalty term, which is needed to offset the extra term in computing the skewness. The consistency of $\hat{S}_{k, T_{ri}}$ is established and presented in the following Proposition.

¹³Note that this same bias is also present if one applies GMM to a regression of p_t against y_t so as to examine the economic value of p_t .

Proposition 1 Under Assumptions A and B, $\hat{S}_k, T_{ri} \rightarrow p S_{k,i}$ as $T_i \rightarrow \infty$ for $i = 0, 1$, where

$S_{k,i} = \lim_{T_i \rightarrow \infty} S_{k,T_i}$ with

$$S_{k,T_i} = \frac{1}{T_i} \sum_{t=1}^{T_i} E \hat{p}_{it}^* + \frac{3}{T_i} \sum_{s=1}^{T_i-1} \sum_{t=s+1}^{T_i} E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-s}^* \right) + \frac{3}{T_i} \sum_{s=1}^{T_i-1} \sum_{t=s+1}^{T_i} E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-s}^2 \right) + \frac{6}{T_i} \sum_{\tau=1}^{T_i-2} \sum_{s=\tau+1}^{T_i-1} \sum_{t=s+1}^{T_i} E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^* \right).$$

The proof of Proposition 1 is shown in the Appendix, and is similar to the proof of the HAC estimator of the long run variance from Newey & West (1987) and Newey & West (1994).

Given the consistent estimator for the skewness of $T_i^{1/3} \bar{\hat{p}}_i$, we are now ready to propose our testing procedures, which, based on Johnson's and Hall's transformations respectively, take into account the skewness of the data in the presence of serial correlation:

$$z_{T_1, T_0}^J = \frac{\left(\bar{\hat{p}}_1 - \bar{\hat{p}}_0 \right) + \frac{\hat{S}_k^D}{6s^2} + \frac{\hat{S}_k^D}{3s^4} \left(\bar{\hat{p}}_1 - \bar{\hat{p}}_0 \right)^2}{s},$$

and

$$z_{T_1, T_0}^H = \frac{\left(\bar{\hat{p}}_1 - \bar{\hat{p}}_0 \right) + \frac{\hat{S}_k^D}{6s^2} + \frac{\hat{S}_k^D}{3s^4} \left(\bar{\hat{p}}_1 - \bar{\hat{p}}_0 \right)^2 + \frac{\left(\hat{S}_k^D \right)^2}{27s^8} \left(\bar{\hat{p}}_1 - \bar{\hat{p}}_0 \right)^3}{s},$$

where $\hat{S}_k^D = \frac{\hat{S}_{k,T_{r1}}}{T_1^2} - \frac{\hat{S}_{k,T_{r0}}}{T_0^2}$. Note that z_{T_1, T_0}^J and z_{T_1, T_0}^H can be rewritten as

$$z_{T_1, T_0}^J = z_{T_1, T_0} + \frac{\hat{S}_k^D}{6s^3} + \frac{\hat{S}_k^D}{3s^5} \left(\bar{\hat{p}}_1 - \bar{\hat{p}}_0 \right)^2. \quad (7)$$

and

$$z_{T_1, T_0}^H = z_{T_1, T_0} + \frac{\hat{S}_k^D}{6s^3} + \frac{\hat{S}_k^D}{3s^5} \left(\bar{\hat{p}}_1 - \bar{\hat{p}}_0 \right)^2 + \frac{\left(\hat{S}_k^D \right)^2}{27s^9} \left(\bar{\hat{p}}_1 - \bar{\hat{p}}_0 \right). \quad (8)$$

Under the null hypothesis of no economic value, both the z_{T_1, T_0}^J and z_{T_1, T_0}^H statistics follow a standard normal distribution as $T_1 \rightarrow \infty$, $T_0 \rightarrow \infty$ and $\frac{T_0}{T_1} \rightarrow c$, such that $0 < c < \infty$. The consistency of $\hat{S}_{k,T_{ri}}$ and s_i^2 , together with the restriction on the relative growth of the sample size, guarantees that the terms other than the first (z_{T_1, T_0}) in Eqs. (7) or (8) are of order $o_p(1)$. Clearly, the z_{T_1, T_0}^J and z_{T_1, T_0}^H tests, like the PT test, are designed specifically to handle serially correlated data. They are complementary rather than alternative procedures to the PT test, in the sense that they deal with different data structures: the procedures proposed here deal with continuous probabilities, while the PT test uses binary variables.

4. Empirical exercises and results

4.1. Testing the value of the forecasts

With the aim of constructing a better series of combined forecasts based only on the valuable individual forecasts, we test the value of the forecasts for each individual forecaster who satisfies the participation requirement. The forecasts for each horizon and subsample from each individual are tested separately from the rest. For horizons $h = 0$ to 4, for a given level of significance, the numbers of forecasters whose forecasts are considered to have value in the sense of Merton (1981) are given in Table 2.¹⁴ Depending on the subsample, at a 1% significance level, about 75% of the forecasters make valuable forecasts for the current quarter. It is clear that the number of forecasters making valuable forecasts decreases sharply after two quarters (i.e., the current and the next quarter). After the third quarter, only a few forecasters are making valuable forecasts. A significantly larger proportion of forecasters make valuable forecasts in subsample 2 than in subsample 1. Surprisingly, beyond two quarters, less than 10% of the forecasters make valuable forecasts for subsample 1, while around 20% of the forecasters still make valuable forecasts for subsample 2. This could be due to the later period having much less volatility in real GDP, except for the short period around the 2008 recession. Note that the aforementioned observations are made even though almost all the individual series look reasonably good when measured by QPS, as is shown in Table 1.¹⁵ Our results are consistent with what has been reported in the literature about the performance of the SPF forecasts (for example, Stark 2010, Lahiri & Wang 2013).

4.2. Beta-transformed linear pool

It is widely accepted that combining forecasts leads to higher accuracy. With a rich data set containing an abundance of individual series of forecasts like the SPF, forecast combination should prove to be helpful. In fact, many studies have shown that a significant performance boost can be obtained by combining survey forecasts properly.¹⁶ However, Ranjan & Gneiting (2010) show that any non-trivially weighted average of two or more distinct and calibrated probability forecasts must be uncalibrated and lack sharpness (i.e., have a low resolution). They propose to apply a beta-transformation to the linearly combined probability forecasts in order to restore calibration and resolution. Let p_{it} be the probability forecast made by forecaster i for time t . The beta-transformed combined forecast p_t takes the form $p_t = H_{\alpha, \beta}(\sum_i w_i p_{it})$, where $H_{\alpha, \beta}(p) = B(\alpha, \beta)^{-1} \int_0^p t^{\alpha-1} (1-t)^{\beta-1} dt$ for $p \in [0, 1]$. $\alpha > 0$ and $\beta > 0$ are the shape parameters for the cumulative distribution function of the beta transformation. The log-likelihood function can then be written as $l = \sum_t y_t \ln[H_{\alpha, \beta}(\sum_i w_i p_{it})] + \sum_t (1 - y_t) \ln[1 - H_{\alpha, \beta}(\sum_i w_i p_{it})]$, where y_t is the binary variable to be forecast Ranjan & Gneiting (2010)

¹⁴We report only the results from the z^H test, since the results from the z^J test are quite similar. The omitted results are available from the authors on request.

¹⁵Following Seillier-Moiseiwitsch & Dawid (1993), we also test whether the forecasts are perfect, in the sense that QPS is not unreasonably larger than its expected value of $\sum f_t (1 - f_t)$, using the following test statistic

$$Y_n = \frac{\sum_{t=1}^T (1-2f_t)(x_t - f_t)}{[\sum_{t=1}^T (1-2f_t)^2 (1-f_t)f_t]^{1/2}} \rightarrow_d N(0, 1)$$

We find that everyone passes the test except for a very few forecasters at long horizons. This apparent lack of power may be due to the fact that the test does not explicitly accommodate serial correlation and skewness.

¹⁶For recent examples using the ECB and the U.S. SPF data, see Genre et al. (2010) and Lahiri et al. (2012) respectively. Clements (2011) studies the properties of various combination schemes for a number of plausible data generating processes.

suggest estimating the parameters of the beta-transformation along with the weights for linear combination by maximum like-lihood. Depending on the restrictions imposed on the parameters of the beta-transformation, three different combination schemes can be obtained. The optimally weighted linear opinion pool (OLP) imposes the restriction $\alpha = \beta = 1$, which is equivalent to not using a beta-transformation but using the linearly combined forecasts directly. The beta-transformed linear opinion pool (BLP) imposes the restriction $\alpha = \beta = 1$, resulting in a transformation that tunes up a forecast if it is bigger than 0.5 and tunes it down otherwise. The asymmetric beta-transformed linear opinion pool (BLPa) is the most general of the three. It only requires $\alpha > 0$ and $\beta > 0$. The equally weighted linear pool is denoted by ELP, and is included as a benchmark.

We follow Ranjan & Gneiting (2010) in combining the valuable individual forecasts identified by the test. However, persistence and distinctness of the individual forecasts and a fully balanced panel are needed in order for the procedure to perform well in practice. To address this issue, we resort to the methodology proposed by Aiolfi & Timmermann (2006). Specifically, we first rank and sort all valuable individual forecasts (as indicated by the test results at the 1% significance level) into four groups, based on their p -values, with group 1 being the group with the lowest p -values and group 4 being the group with the highest p -values. Then, we pool the individual forecasts in each group linearly using the ELP method. The group forecasts constructed this way are distinct in performance and have fewer missing values than the individual forecasts. Since the forecasters in our sample do not move across groups often, the group forecasts are also more persistent than the individual forecasts. However, there is still a small number of missing values in the group forecasts (around 10%). We impute these by using the simple average of the available group forecasts (from other groups) for the same period. This gives us a fully balanced panel of group forecasts. Since the number of individual forecasters making valuable four-quarter-ahead forecasts is small, we restrict the combination to the first three quarters only. This in-sample forecast combination exercise is conducted separately for each horizon and subsample. We consider an out-of-sample counterpart of this exercise in the last subsection.

4.3. Forecast combination and evaluation

Table 3 gives the maximum likelihood estimates of the weights and the parameters for the beta-transformation for each series of group forecasts. The top panel is for the in-sample combination exercises discussed in this subsection, and the bottom panel is for the out-of-sample combination exercises presented in the following subsection. For each horizon, the log-likelihood value increases as we go from OLP to BLP to BLPa, as the restrictions α on and β become more relaxed. In order to give a thorough evaluation of the performance of these forecasts, we examine additional measures of performance. More specifically, we examine the hit rate, the false alarm rate, the odds ratio, the Kuiper score, the QPS, the calibration error and the resolution of the forecasts. Let the total number of forecasts from a forecaster/group be $N = N_1 + N_0$. The hit rate (P_1) is the mean of all N_1 forecast probabilities conditioned on a real GDP decline. The false alarm rate (P_0) is the mean of all N_0 forecast probabilities conditioned on no real GDP decline. Good forecasts have high hit rates and low false alarm rates. The odds ratio is defined as $(P_1/(1 - P_1))/(P_0/(1 - P_0))$. Higher odds ratios are more desirable and pertinent, particularly for rare events, see Stephenson (2000). The

calibration error and the resolution measures are calculated by decomposing the QPS. According to Brier(1950), Murphy (1973), and Dawid (1986), we decompose the QPS as follows:¹⁷

$$E[P(\hat{y}_t=1) - y_t]^2 = E[y_t - E(y_t)]^2 + E[P(\hat{y}_t=1) - E[y_t|P(\hat{y}_t=1)]]^2 - E[E[y_t|P(\hat{y}_t=1)] - E(y_t)]^2.$$

The first term on the right hand side is the variance of the outcome series, which is independent of the forecasts. The second term is the mean-squared calibration error, which takes a value between 0 and 1, with 0 being perfect calibration and 1 representing the opposite. The third term measures the resolution or of the forecasts, with a minimum value of 0 indicating no resolution.¹⁸

Table 4 reports these performance measures. They are based on the forecasts from both subsamples, with the second subsample appended to the end of the first subsample. Here, we evaluate the four group forecasts, the combined forecasts under different restrictions on the beta-transformation, and the benchmark forecasts, i.e., the simple average of all individual forecasts (overall mean) in the original SPF data without the participation requirement imposed. In general, the group forecasts outperform the average individual forecasts,¹⁹ and the combined forecasts outperform both the group and the average individual forecasts. Comparing the four group forecasts across the different performance measures, we find that a low p -value from the test does not imply a better performance across all of the measures, which is hardly surprising. In fact, we not find any single measure to be sufficient in determining the quality of the forecasts. Looking at the performance of the combined forecasts, we find that even though the beta-transformation reduces the calibration error only modestly, the combined forecasts produced by the BLP and the BLPa methods do have much higher odds ratios than those produced by the OLP or ELP methods. This is especially true for longer horizon forecasts. For two-quarter-ahead the BLPa combined forecasts have an odds ratio of 6.0, which is almost double that of the ELP forecasts, at 3.1. Moreover, the beta-transformed combined forecasts seem to have a better performance than the average forecasts of all the individuals at longer horizons. The BLPa combined two-quarter-ahead forecasts performance best according to six of the seven measures. In addition, in terms of the odds ratio, the Kuiper score, and the QPS, the BLPa two-quarter-ahead combined forecasts are better than all four group means.

We test for calibration of the simple average and BLPa combined forecasts using the prequential test of Seillier-Moiseiwitsch & Dawid (1993) (SM-D). Lahiri & Wang (2013) report the results of this test using the overall average probabilities. Since many of the original probability intervals contain very few observations, following SM-D's guideline as far as ensuring an adequate number of observations in each interval, we consolidate the eleven probability bins into only three groups, and compute the calibration test statistic for

¹⁷For a recent discussion of this decomposition and some alternatives, see Lahiri & Wang (2013). For a recent application using forecasts for British GDP and Galbraith & van Norden (2012).

¹⁸The resolution is sometimes reported after dividing it by the variance of the outcome series, so that it also ranges from 0 to 1, with perfect forecasts having a resolution of 1. However, we report the "raw" number without this normalization here, as our exercises concern only one outcome series.

¹⁹Statistics for individual forecasts for each subsample separately are omitted for brevity, are available from the authors.

each forecast horizon. The calculated χ^2 values are such that the linear opinion pool forecasts for all of the quarters are found to be uncalibrated at the usual 5% level of significance with 2 degrees of freedom. However, as designed by Ranjan & Gneiting (2010), the BLPa forecasts are found to be calibrated. Unfortunately, due to the sparseness of observations in the higher probability bins, we cannot make any inference on calibration at higher deciles of forecast probabilities. Our test are reported in Table 5.

4.4. Additional graphical diagnostics

As has been described, Figures 1a and 1b show ranked individual KSS scores for three groups of forecasters, separately for each horizon and subsample. In each figure, the top plot shows the KSS of all forecasters who satisfy our participation requirement; the middle plot shows the same for the forecasters who make valuable forecasts, as determined by our test at the 1% significance level; and the bottom plot shows the scores of the remaining forecasters, who are not considered to have made valuable forecasts. It is obvious from the figures that, in a horizon-by-horizon comparison, most forecasters whose forecasts are considered valuable have higher KSSs than those whose forecasts are not valuable, although there are some exceptions. As is shown in the middle plot of Figure 1b, a few of the forecasters are identified not to have made valuable forecasts in spite of having relatively high KSS scores, especially for current quarter forecasts. This could be due to the underlying autocorrelation and skewness of the forecasts. Comparing the forecasts within each plot, we see that the forecast performance deteriorates rapidly from the current quarter to one quarter ahead, then slowly deteriorates further as the horizon lengthens. On the other hand, the middle plot of Figure 1b shows that three forecasters are considered to have made valuable forecasts for the current and the next quarter, in spite of their low KSSs. This is another example of the need for serial correlation and skewness correction in detecting the value of forecasts. Of the forecasters who make valuable forecasts, more than half of them have KSS scores for two-quarter-ahead forecasts which are similar to or lower than those for forecasts made three or four quarters ahead. This indicates that the forecasts start to lose value at two quarters ahead, even though they are still considered valuable statistically. Very few forecasters with KSS scores near zero are identified to have made valuable forecasts, at least for the first two horizons. From these results, which are largely consistent with the previous observations made based on the quantitative measures, the role of the Kuiper score in determining the practical value of the forecasts becomes clear. In general, a low Kuiper score usually indicates a poor performance, while not all forecasts with high Kuiper scores are valuable, presumably due to their associated additional variance, serial correlation and/or skewness.

We perform an additional graphical diagnostic check which is shown in two time series plots in Figure 3. The top panel of Figure 3 shows the comparison for one-quarter-ahead forecasts, where the mean of the individual forecasts from the best group is compared against the simple average benchmark. Overall, the two sets of forecasts are similar graphically. The mean of the best group correctly identified two periods of real GDP declines that the benchmark forecasts failed to identify, around 1978 and 2001. The bottom panel of Figure 3 compares the two-quarter-ahead BLPa combined forecasts with the benchmark mean forecasts of all individuals. Here, a clear advantage of the combined forecasts is observed. In the later years, the BLPa combined forecasts show an almost

uniformly better performance, i.e., the forecasts are high during periods with real GDP declines and low during non-decline periods, except for a few quarters around mid-2001. In the earlier years, except perhaps for the 1978 period, the BLPa combined forecasts also perform better than the benchmark.

These observations echo the results presented in Table 4 and confirm the reasonable quality of the simple average benchmark forecasts. However, more importantly, they also show the superiority of the beta-transformed optimally-weighted forecasts, which are based on the forecasters who make valuable forecasts, as identified by the test. We note that it may seem from the number of forecasters making valuable forecasts (as shown in Table 2 and the graphical checks, especially Figure 3) as though most forecasters perform better in general in the later years, covered by subsample 2. Using a fixed benchmark forecast that matches the overall sample proportion of periods with a real GDP decline, we can compute the skill scores for each of the subsample periods. The results show that, except for the current quarter forecasts (where they turn out to be almost the same), the forecasts made during the later years do indeed display higher skill levels.

4.5. An out-of-sample evaluation

As was shown in the previous subsection, the improvement in the forecasting performance which is attributed to combining only the valuable forecasts, as identified by the test, demonstrates its usefulness. While the purpose of the test is merely to assess a given set of forecasts and determine its value, it would be more useful if the test could help with out-of-sample forecast combination. To explore this possibility, we adapt the procedure used in the in-sample forecast combination exercise in the previous section for performing out-of-sample combination. Specifically, we use the part of the first subsample from 1969:I to 1978:IV as the training sample, and test each forecaster's performance during this period.²⁰ The forecasters that are found to have made valuable forecasts during the training sample period are retained for combination using the evaluation sample from 1979:I to 1990:I. The combined forecasts for this out-of-sample period are then evaluated using the same set of performance measures as were used in the previous exercise.

Before we discuss the results, several remarks on the setup of this exercise in order. First, in order to evaluate the value of the forecasts properly using test, the track record of a forecaster must cover sufficiently long periods of both real GDP declines and real GDP growth. As there are only three episodes of real GDP declines in the second subsample, the issue of missing data means that not enough observations are available for most forecasters if the subsample is further divided into a training sample and an evaluation sample. Therefore, in our out-of-sample exercise, only subsample 1 is considered. Second, it is computationally feasible to apply the test to either an expanding window of forecasts or a rolling window of forecasts, thus allowing the testing and forecast combination procedure to be carried out in real time. However, for the majority of the periods in which there is no real GDP decline, adding one more observation to the training sample does not alter the test result. This dissuades us from attempting a real-time combination and evaluation, since the

²⁰The training sample is selected so as to leave enough observations of both real GDP declines and real GDP growth in both the training and evaluation samples.

conclusions we would obtain from a real time implementation could also be obtained from a simple out-of-sample exercise. Third, the crucial element of a successful combination of forecasts is that a forecaster's past performance is a good predictor of her future performance. Obviously, this is a data-specific property and is not related to forecast combination procedures. Therefore, one should avoid interpreting the performance of the combined forecasts as the performance of the test, even though, in general, a more powerful test should help to produce better combined forecasts.

The estimated combination weights are reported in the bottom panel of Table 3, and the performance measures are reported in Table 6. For current-quarter forecasts, the ELP and OLP combined forecasts outperform the overall mean benchmark in terms of most of the performance measures. This indicates the ability of the test to identify good forecasters out-of-sample, based on their track record. However, for longer horizon forecasts, even though the combined forecasts are marginally better than the benchmark, neither the combined forecasts nor the benchmark appear to display any skill. This suggests the basic unpredictability of the target variable at longer horizons. This could also be due to fact that the evaluation sample period is very volatile, with frequent real GDP declines that are difficult to foresee. We again note that, since the predictability of forecasters' performances is a key ingredient of successful forecast combination, it is unlikely that the results reported here can be generalized readily other variables and data sets, even though the empirical strategy and procedure used here should be of value more generally.

5. Concluding remarks

In order to identify the forecasters whose forecasts have “no value” in a KSS sense (or in the sense of Merton (1981)), we propose the use of a serial correlation and skewness-robust Welch-type test that utilizes the probability forecasts directly, without requiring the forecasts to be converted to binary forecasts before testing. We then test, combine, and analyze in detail the probability forecasts of a real GDP decline using data from the U.S. Survey of Professional Forecasters from 1968Q4 to 2012Q1. While the data set spans more than four decades, covering many recessions and cyclical downturns, it contains a large number of missing values. In order to mitigate this problem, we follow the multi-step methodology suggested by Aiolfi & Timmermann (2006), by first sorting the frequent forecasters into four clusters based on their past performances, then applying the BLP methodology of Ranjan & Gneiting (2010) to these clusters in order to obtain calibrated combined forecasts. We compare the combined forecasts based on the valuable individual forecasts to the simple average (ELP) of all the forecasts for current-, one-, and two-quarter horizons.

Our findings suggest that the overall value of the SPF probability forecasts of real GDP decline diminishes sharply and becomes low after two quarters, confirming previous similar findings in the literature. This also means that an indiscriminate use of the “raw” individual forecasts and the equally weighted average forecasts may have little practical value when the forecast horizon extends beyond two quarters. Our test results suggest that, among the 50 or so forecasters who forecast frequently during both periods of real GDP declines and periods of positive growth, only half are making valuable forecasts, at 1% significance level, for

short horizons. In addition, we find that the performances of individual forecasts are relatively better in the second subsample than in the first subsample.

In order to induce persistence and distinctness in performance and to mitigate the issues caused by the large amount of missing data in the SPF forecasts, we group the series of valuable individual forecasts into four groups based on their p -values from the test, and then combine them using the beta-transformed linear pool (BLP). After examining the performances of the benchmark, individual groups, and the combined forecasts, we conclude that the BLP combined forecasts and the equally-weighted forecasts of the top forecasters outperform the simple average based on all individual forecasters. We also compare the performances of the combined and the benchmark forecasts using a number of measures, including calibration and resolution, and find that the ELP combined forecasts based on valuable forecasts outperform the simple average benchmark based on all forecasts. In addition, our results show that, for two-quarter-ahead forecasts, combining the grouped forecasts with a beta-transformed linear pool greatly enhances the performance, while the corresponding simple average benchmark forecasts have little practical value at this horizon and beyond.

Based on the results obtained in this study, we suggest that the following procedure be used when attempting to combine these subjective probability forecasts in real time. In each period, first exclude from consideration the forecasters that have a very limited track record of participation. Then, rank the individual forecasters who participated in that period according to their Kuiper scores based on their historical performances. For those with reasonably high Kuiper scores, test the value of the recorded forecasts using our test procedure, and exclude anyone who fails the test at a pre-assigned significance level. Finally, combine the forecasts of the remaining forecasters, i.e., those passing the test, using the BLP or BLPa methodology with weights estimated using historical data.

While our results suggest that testing the individual forecasters and excluding those who do not have economic value helps to identify good forecasters *ex ante*, the performance of the combined forecasts based on these forecasters may not always be satisfactory, especially at long horizons. This is due to the fact that the *ex ante* forecast performance of individual forecasters is highly data-dependent, and is unrelated to the empirical method that we choose to compare them. Nevertheless, the test proposed in this paper provides an additional dimension along which the economic value of potentially serially correlated and skewed forecasts can be assessed, independent of any threshold parameter that would be required to convert the probabilities to binary forecasts.

Acknowledgements

The authors thank, without implicating, the anonymous referee and Prakash Loungani for their helpful comments.

Appendix: Technical proofs

Definition:

$$S_{k,T_{ri}} = \frac{1}{T_i} \sum_{t=1}^{T_i} E \hat{p}_{it}^{*3} + \frac{3}{T_i} \sum_{s=1}^{T_{ri}} \sum_{t=s+1}^{T_i} E \left(\hat{p}_{i,t}^{*2} \hat{p}_{i,t-s}^* \right) + \frac{3}{T_i} \sum_{s=1}^{T_{ri}} \sum_{t=s+1}^{T_i} E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-s}^{*2} \right) + \frac{6}{T_i} \sum_{\tau=1}^{T_{ri}} \sum_{s=\tau+1}^{T_i} \sum_{t=s+1}^{T_i} E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^* \right)$$

$$\tilde{S}_{k,T_{ri}} = \frac{1}{T_i} \sum_{t=1}^{T_i} \hat{p}_{it}^{*3} + \frac{3}{T_i} \sum_{s=1}^{T_{ri}} \omega(s, T_{ri}) \sum_{t=s+1}^{T_i} \hat{p}_{i,t}^{*2} \hat{p}_{i,t-s}^* + \frac{3}{T_i} \sum_{s=1}^{T_{ri}} \omega(s, T_{ri}) \sum_{t=s+1}^{T_i} \hat{p}_{i,t}^* \hat{p}_{i,t-s}^{*2} + \frac{6}{T_i} \sum_{\tau=1}^{T_{ri}} \omega(\tau, T_{ri}) \sum_{s=\tau+1}^{T_i} \omega(s, T_{ri}) \hat{p}_{i,t}^* \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^*.$$

Lemma 1 Under Assumption A, $\sqrt{T_i} (\bar{\hat{p}} - p_i) \rightarrow_p N(0, \sigma_i^2)$ as $T_i \rightarrow \infty$

Proof. See Corollary 5.3 (ii) in Hall & Heyde (1980).

Lemma 2 Suppose Assumptions A and B hold. Then $S_{k,i} = \lim_{T_i \rightarrow \infty} S_{k,T_i} < \infty$ and $S_{k,T_{ri}} - S_{k,T_i} \rightarrow 0$ as $T_i \rightarrow \infty$.

Proof. By definition

$$S_{k,T_i} = \frac{1}{T_i} \sum_{t=1}^{T_i} E \hat{p}_{it}^{*3} + \frac{3}{T_i} \sum_{s=1}^{T_i} \sum_{t=s+1}^{T_i} E \left(\hat{p}_{i,t}^{*2} \hat{p}_{i,t-s}^* \right) + \frac{3}{T_i} \sum_{s=1}^{T_i-1} \sum_{t=s+1}^{T_i} E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-s}^{*2} \right) + \frac{6}{T_i} \sum_{\tau=1}^{T_i-2} \sum_{s=\tau+1}^{T_i-1} \sum_{t=s+1}^{T_i} E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^* \right),$$

since $\hat{p}_{i,t}^* = \hat{p}_{i,t} - p_i$ and $p_{i,t}$ is stationary with $E p_{i,t} = p_i$ and $E |p_{i,t}|^{8+\delta} < \infty$ for some $\delta > 0$.

It follows immediately that $\frac{1}{T_i} \sum_{t=1}^{T_i} E \hat{p}_{it}^{*3} = E \hat{p}_{i,t}^{*3} < \infty$. Next, by using the fact that $E \hat{p}_{i,t-s}^* = 0$ and Corollary 6.16 of White (1984), we obtain

$$E \left(\hat{p}_{i,t}^{*2} \hat{p}_{i,t-s}^* \right) = E \left(\left(\hat{p}_{i,t}^{*2} - E \hat{p}_{i,t}^{*2} \right) \hat{p}_{i,t-s}^* \right) \leq [\alpha_i(s)]^{\frac{\delta}{2+2\delta}} M$$

for some positive finite M . Hence,

$$\begin{aligned} \frac{3}{T_i} \sum_{s=1}^{T_i-1} \sum_{t=s+1}^{T_i} \left| E \left(\hat{p}_{i,t}^{*2} \hat{p}_{i,t-s}^* \right) \right| &\leq \frac{3}{T_i} \sum_{s=1}^{T_i-1} \sum_{t=s+1}^{T_i} [\alpha_i(s)]^{\frac{\delta}{2+2\delta}} M \\ &= 3M \sum_{s=1}^{T_i-1} [\alpha_i(s)]^{\frac{\delta}{2+2\delta}} \left(\frac{T_i-s}{T_i} \right) \\ &\leq 3M \sum_{s=1}^{T_i-1} [\alpha_i(s)]^{\frac{\delta}{2+2\delta}} \\ &< \infty, \end{aligned}$$

because $\alpha_i(s)$ is of size $(2+2\delta)/\delta$. Similarly, we have

$$\frac{3}{T_i} \sum_{s=1}^{T_i-1} \sum_{t=s+1}^{T_i} \left| E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-s}^{*2} \right) \right| < \infty.$$

Turning to the term $\frac{6}{T_i} \sum_{\tau=1}^{T_i-2} \sum_{s=\tau+1}^{T_i-1} \sum_{t=s+1}^{T_i} E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^* \right)$ where $0 < \tau < s$, it can be seen that

$$\begin{aligned}
\frac{6}{T_i} \sum_{\tau=1}^{T_i-2} \sum_{s=\tau+1}^{T_i-1} \sum_{t=s+1}^{T_i} E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^* \right) &\leq 6M \sum_{\tau=1}^{T_i-2} \sum_{s=\tau+1}^{T_i-1} [\alpha_i(s) \alpha_i(\tau)]^{\frac{\delta}{2+2\delta}} \left(\frac{T_i-s}{T_i} \right) \\
&< 6M \sum_{\tau=1}^{T_i-2} [\alpha_i(\tau)]^{\frac{\delta}{2+2\delta}} \sum_{s=1}^{T_i-1} [\alpha_i(s)]^{\frac{\delta}{2+2\delta}} \\
&< \infty.
\end{aligned}$$

Thus, S_{k,T_i} is bounded above for all T_i , So $\lim_{T_i \rightarrow \infty} S_{k,T_i}$ exists and $\lim_{T_i \rightarrow \infty} S_{k,T_i} < \infty$, $S_{k,T_i} < \infty$.
The proof of $S_{k,T_{ri}} - S_{k,T_i} \rightarrow 0$ as $T_i \rightarrow \infty$ parallels exactly the proof of Lemma 6.17 of White (1984) and hence is omitted here.

Proposition 1 Under Assumption A and B, $\hat{S}_{k,T_{ri}} - S_{k,T_i} \rightarrow_p 0$ as $T_i \rightarrow \infty$.

Proof. By Lemma 1,

$$\tilde{p}_{i,t} = \hat{p}_{i,t}^* + (\tilde{p}_i - p_i) = \hat{p}_{i,t}^* + O_p \left(\frac{1}{T_i^{1/2}} \right), \quad (9)$$

which, along with the stationarity of $\hat{p}_{i,t}^*$, implies that $\hat{S}_{k,T_{ri}} = \tilde{S}_{k,T_{ri}} + O_p \left(\frac{1}{T_i^{1/2}} \right)$. Since $S_{k,T_i} - \tilde{S}_{k,T_{ri}} \rightarrow 0$ as $T_i \rightarrow \infty$ and $S_{k,i} = \lim_{T_i \rightarrow \infty} S_{k,T_i}$, by virtue of Lemma 2, it suffices to show $\tilde{S}_{k,T_{ri}} - S_{k,T_{ri}} \rightarrow_p 0$ as $T_i \rightarrow \infty$. By triangle inequality.

$$\begin{aligned}
\left| \tilde{S}_{k,T_{ri}} - S_{k,T_{ri}} \right| &\leq \left| \frac{1}{T_i} \sum_{t=1}^{T_i} \left(\hat{p}_{it}^{*3} - E \hat{p}_{it}^{*3} \right) \right| + \frac{3}{T_i} \sum_{s=1}^{T_{ri}} |\omega(s, T_{ri}) - 1| \sum_{t=s+1}^{T_i} \left| E \left(\hat{p}_{i,t}^{*2} \hat{p}_{i,t-s}^* \right) \right| \\
&\quad + \frac{3}{T_i} \sum_{s=1}^{T_{ri}} |\omega(s, T_{ri}) - 1| \sum_{t=s+1}^{T_i} \left| E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-s}^{*2} \right) \right| \\
&\quad + \frac{6}{T_i} \sum_{r=1}^{T_{ri}} \omega(\tau, T_{ri}) \sum_{s=\tau+1}^{T_{ri}} |\omega(s, T_{ri}) - 1| \sum_{t=s+1}^{T_i} \left| E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^* \right) \right| \\
&\quad + \frac{6}{T_i} \sum_{r=1}^{T_{ri}} |\omega(\tau, T_{ri}) - 1| \sum_{s=\tau+1}^{T_{ri}} \omega(s, T_{ri}) \sum_{t=s+1}^{T_i} \left| E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^* \right) \right| \\
&\quad + \frac{6}{T_i} \sum_{r=1}^{T_{ri}} |\omega(\tau, T_{ri}) - 1| \sum_{s=\tau+1}^{T_{ri}} |\omega(s, T_{ri}) - 1| \sum_{t=s+1}^{T_i} \left| E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^* \right) \right| \\
&\quad + \left| \frac{3}{T_i} \sum_{s=1}^{T_{ri}} \omega(s, T_{ri}) \sum_{t=s+1}^{T_{ri}} \left(\hat{p}_{i,t}^{*2} \hat{p}_{i,t-s}^* - E \left(\hat{p}_{i,t}^{*2} \hat{p}_{i,t-s}^* \right) \right) \right| \\
&\quad + \left| \frac{3}{T_i} \sum_{s=1}^{T_{ri}-1} \omega(s, T_{ri}) \sum_{t=s+1}^{T_{ri}} \left(\hat{p}_{i,t}^* \hat{p}_{i,t-s}^{*2} - E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-s}^{*2} \right) \right) \right| \\
&\quad + \left| \frac{6}{T_i} \sum_{\tau=1}^{T_{ri}} \omega(\tau, T_{ri}) \sum_{s=\tau+1}^{T_{ri}} \omega(s, T_{ri}) \sum_{t=s+1}^{T_i} \left[\hat{p}_{i,t}^* \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^* - E \left(\hat{p}_{i,t}^* \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^* \right) \right] \right|.
\end{aligned}$$

Since $\hat{p}_{i,t}$ is stationary and has $\alpha_i(t)$ of size $(2+2\delta)/\delta$ and $E \hat{p}_{i,t}^{4+\delta} < \infty < \infty$ for some $\delta > 0$, by Theorem 2.10 of Mcleish (1975), we have

$$\frac{1}{T_i} \sum_{t=1}^{T_i} E \hat{p}_{i,t}^{*3} \rightarrow_p E \hat{p}_{i,t}^{*3} < \infty.$$

Since $p_{i,t}$ is stationary and has $\alpha_i(t)$ of size $(2 + 2\delta)/\delta$ and $E\hat{p}_{i,t}^{4+\delta} < \infty < \infty$ for some $\delta > 0$, by Theorem 2.10 of Mcleish (1975), we have

$$\frac{1}{T_i} \sum_{t=1}^{T_i} E\hat{p}_{i,t}^{*3} \rightarrow_p E\hat{p}_{i,t}^{*3} < \infty.$$

Since $p_{i,t}$ is stationary and has $\alpha_i(t)$ of size $(2 + 2\delta)/\delta$ and $E\hat{p}_{i,t}^{8+\delta} < \infty < \infty$ for some $\delta > 0$, by Theorem 2.10 of Mcleish (1975), the first term goes to zero. Next, as was shown in Lemma

2, we have $|E(\hat{p}_{i,t}^{*2} \hat{p}_{i,t-s}^*)| \leq [\alpha_i(s)]^{\frac{\delta}{2+2\delta}} M$ for some finite constant M . Hence

$|E(\hat{p}_{i,t}^{*2} \hat{p}_{i,t-s}^*)| \leq [\alpha_i(s)]^{\frac{\delta}{2+2\delta}} M$ for all T_i and s . Since $\omega(s, T_{ri}) = 1 - \frac{s}{T_{ri}+1} \rightarrow 1$ for each s , the second term tends to zero as $T_i \rightarrow 0$ by virtue of the dominated convergence theorem.

By similar arguments, the third, fourth, fifth and sixth terms all approach zero as $T_i \rightarrow 0$.

Next, let $z_{t\tau s} = \hat{p}_{i,t} \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^* - E(\hat{p}_{i,t} \hat{p}_{i,t-\tau}^* \hat{p}_{i,t-s}^*)$; then, it follows immediately from

Assumption A that $E s_{t\tau s}^{2+2\delta} \leq D^* < \infty$ for some constant D^* and $\delta > 0$. Then, using an argument similar to that in Theorem 2 of Newey & West (1987), we obtain

$$E \left[\sum_{t=s+1}^{T_i} E z_{t\tau s} \right]^2 \leq (T_i - s)(s+1) \Delta \leq T_i (s+1) \Delta$$

for $s > 0$ and some $\Delta < \infty$. Hence,

$$\begin{aligned} & P \left(\left| \frac{1}{T_i} \sum_{\tau=1}^{T_{ri}} \omega(\tau, T_{ri}) \sum_{s=\tau+1}^{T_{ri}} \omega(s, T_{ri}) \sum_{t=s+1}^{T_i} E z_{t\tau s} \right| > \varepsilon \right) \\ & \geq P \left(\left| \frac{1}{T_i} \sum_{\tau=1}^{T_{ri}} \omega(\tau, T_{ri}) \sum_{s=\tau+1}^{T_{ri}} \omega(s, T_{ri}) \sum_{t=s+1}^{T_i} E z_{t\tau s} \right| > \varepsilon \right) \\ & \leq \sum_{\tau=1}^{T_{ri}} \sum_{s=\tau+1}^{T_{ri}} P \left(\left| \frac{1}{T_i} \sum_{t=s+1}^{T_i} E z_{t\tau s} \right| > \frac{\varepsilon}{2T_{ri}(T_{ri}-1)} \right) \\ & \leq \sum_{\tau=1}^{T_{ri}} \sum_{s=\tau+1}^{T_{ri}} \frac{E \left(\frac{1}{T_i} \sum_{t=s+1}^{T_i} E z_{t\tau s} \right)^2}{\left(\frac{\varepsilon}{2T_{ri}(T_{ri}-1)} \right)^2} \\ & \leq \frac{4T_{ri}^2 (T_{ri}-1)^2}{T_i \varepsilon^2} \Delta \sum_{\tau=1}^{T_{ri}} \sum_{s=1}^{T_{ri}} (s+1) \\ & = \frac{2T_{ri}^4 (T_{ri}-1)^2 (T_{ri}+2)}{T_i \varepsilon^2} \\ & \rightarrow 0. \end{aligned}$$

provided $\frac{T_{ri}^7}{T_i} \rightarrow 0$. That is, the ninth term goes to zero if T_{ri}^7 grows more slowly than T . By a similar argument, the seventh and eighth terms also go to zero under Assumptions A and B. Therefore, we can conclude that

$$\left| \tilde{S}_{k,T_{ri}} - S_{k,T_{ri}} \right| \rightarrow_p 0,$$

which completes the proof.

Biography

Kajal Lahiri is a distinguished professor of economics at the University at Albany-SUNY, a fellow of CESifo Research Network, and an honorary fellow of the International Institute of Forecasters. He received his doctorate in economics from the University of Rochester. Author of numerous books and articles, Dr. Lahiri created the Transportation Services Index, a measure of output in the transportation sector that is reported monthly by the US Department of Transportation. He is on the editorial boards of the *Journal of Econometrics*, the *International Journal of Forecasting*, *Empirical Economics*, and the *Journal of Business Cycle Measurement and Analysis*. His current research deals with economic forecasting, and also with minority health, with the latter research being funded by the National Institutes of Health.

Huaming Peng (Ph.D., Yale University) is Assistant Professor of economics at the Department of Economics of SUNY at Albany. His research interests include theoretical and applied econometrics, and time series forecasting.

Yongchen Zhao is a doctoral candidate at the Department of Economics, the University at Albany-SUNY. His research interests are in the areas of forecasting, applied econometrics, and finance. His recent work has centered on forecasting with survey data and many predictors, including issues in forecast combination, forecast evaluation and testing, as well as applications to the hedge fund industry.

References

- Aiolfi M, Timmermann A. Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*. 2006; 135(1-2):31–53.
- Braun P, Yaniv I. A case study of expert judgment: Economists' probabilities versus base rate model forecasts'. *Journal of Behavioral Decision Making*. 1992; 5:217–231.
- Brier G. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. 1950; 78:1–3.
- Capistran C, Timmermann A. Forecast combination with entry and exit of experts. *Journal of Business and Economic Statistics*. 2009; 27(4):428–440.
- Clements M, Harvey DI. Combining probability forecasts. *International Journal of Forecasting*. 2011; 27(2):208–223.
- Clements MP. Consensus and uncertainty: Using forecast probabilities of output declines. *International Journal of Forecasting*. 2008; 24:76–86.
- Clements MP. Internal consistency of survey respondents' forecasts: Evidence based on the Survey of Professional Forecasters. In: Castle, JL.; Shephard, N.; Hendry, DF.; Castle, J.; Shephard, N., editors. *The methodology and practice of econometrics*. Oxford University Press; Oxford and New York: 2009. p. 206–226.
- Clements MP. An empirical investigation of the effects of rounding on the SPF probabilities of decline and output growth histograms. *Journal of Money, Credit and Banking*. 2011; 43(1):207–220.

- Clements MP, Harvey DI. Forecast encompassing tests and probability forecasts. *Journal of Applied Econometrics*. 2010; 25(6):1028–1062.
- Croushore D. Introducing: The Survey of Professional Forecasters. *Federal Reserve Bank of Philadelphia Business Review*. 1993:3–13.
- Dawid A. Probability forecasting. *Encyclopedia of Statistical Sciences*. 1986
- Engelberg J, Manski C, Williams J. Assessing the temporal variation of macroeconomic forecasts by a panel of changing composition. *Journal of Applied Econometrics*. 2010; 26(7):1059–1078.
- Galbraith JW, van Norden S. Kernel-based calibration diagnostics for recession and International *Journal of Forecasting*. 2011; 27(4):1041–1057.
- Galbraith JW, van Norden S. Assessing gross domestic product and ability forecasts derived from Bank of England fan charts. *Journal of the Royal Statistical Society, Series A(Statistics in Society)*. 2012; 175:713–727.
- Genre V, Kenny G, Meyler A, Timmermann A. Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*. 2010; 27(2):108–121.
- Graham HR. Is a group of forecasters better than one? *Journal of Business*. 1996; 69(2):193–232.
- Granger, CWJ.; Pesaran, MH. A decision theoretic approach to forecast evaluation. In: Chan, W-S.; Li, WK.; Tong, H., editors. *Statistics and finance*. Imperial College Press and Distributed by World Scientific Pub; London and River Edge and NJ: 2000. p. 261–278.
- Guo JH, Luh WM. An invertible transformation two-sample trimmed-statistic under heterogeneity and nonnormality. *Statistics and Probability Letters*. 2000; 49:1–7.
- Hall P. On the removal of skewness by transformation. *Journal of the Royal Statistical Society, Series B*. 1992; 54:221–228.
- Hall, P.; Heyde, C. *Martingale limit theory and its applications*. Academic Press; New York: 1980.
- Henriksson RD, Merton RC. On market timing and investment performance, II: Statistical procedures for evaluating forecasting skills. *Journal of Business*. 1981; 54(4):513–533.
- Johnson NJ. Modified t-tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*. 1978; 73:536–544.
- Lahiri K, Monokroussos G, Zhao Y. The yield spread puzzle and the information content of SPF forecasts. *Economics Letters*. 2013; 118:219–221.
- Lahiri, K.; Peng, H.; Zhao, Y. *Forecast combination in incomplete panels*. Department of Economics Working Paper, University at Albany-SUNY; 2012.
- Lahiri K, Wang JG. Subjective probability forecasts for recessions: Guidelines for use. *Business Economics*. 2006; 41(2):26–37.
- Lahiri K, Wang JG. Evaluating probability forecasts for GDP declines using alter native methodologies. *International Journal of Forecasting*. 2013; 29:175–190.
- Luh WM, Guo JH. A powerful transformation trimmed mean method for one way fixed effects ANOVA model under non-normality and inequality of variances. *British Journal of Mathematical and Statistical Psychology*. 1999; 52:303–320.
- McLeish DL. A maximal in equality and dependent strong laws. *Annals of Probability*. 1975; 3:829–839.
- Merton RC. On market timing and investment performance, I: An equilibrium theory of value for market forecasts. *Journal of Business*. 1981; 54(3):363–406.
- Murphy A. A new vector partition of the probability score. *Journal of Applied Meteorology*. 1973; 12:595–600.
- Newey WK, West KD. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*. 1987; 55:703–708.
- Newey WK, West KD. Automatic lag selection in covariance estimation. *Review of Economic Studies*. 1994; 61:631–654.
- Pesaran MH, Timmermann A. A simple nonparametric test of predictive performance. *Journal of Business and Economic Statistics*. 1992; 10:461–465.
- Pesaran MH, Timmermann A. A generalization of the non-parametric Henriksson-Merton test of market timing. *Economics Letters*. 1994; 44(1–2):1–7.

- Pesaran MH, Timmermann A. Testing dependence among serially correlated multicategory variables. *Journal of the American Statistical Association*. 2009; 104:325–337.
- Ranjan R, Gneiting T. Combining probability forecasts. *Journal of the Royal Statistical Society, Series B(Statistical Methodology)*. 2010; 72(1):71–91.
- Rudebusch GD, Williams JC. Forecasting recessions: The puzzle of the enduring power of the yield curve. *Journal of Business and Economic Statistics*. 2009; 27:492–503.
- Schnader MH, Stekler HO. Evaluating predictions of change. *Journal of Business*. 1990; 63(1):99–107.
- Seillier-Moiseiwitsch F, Dawid AP. On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association*. 1993; 88(421):355–359.
- Sohn KT, Park SM. Guidance on the choice of threshold for binary forecast modeling. *Advances in Atmospheric Sciences*. 2008; 25(1):83–88.
- Stark T. Realistic evaluation of real-time forecasts in the Survey of Professional Forecasters. Federal Reserve Bank of Philadelphia Research Special Report. 2010
- Stekler HO. Are economic forecasts valuable? *Journal of Forecasting*. 1994; 13:495–505.
- Stephenson DB. Use of the odds ratio for diagnosing forecast skill. *Weather and Forecasting*. 2000; 15:221–232.
- Stock JH, Watson MW. Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*. 2003; 41:788–829.
- White, H. Asymptotic theory for econometricians. Academic Press; New York: 1984.

Figure 1a.

Comparison of ranked individual KSS – 1969: I to 1990: I.

For all subfigures, KSS is plotted on the vertical axis, and the number of forecasters is plotted on the horizontal axis. Each point in the graph represents one forecaster. Points joined by a curve represent forecasters forecasting for the same horizon. For each horizon, forecasters are ranked based on KSS. The longer a curve is, i.e., the more points are connected by a curve, the more forecasters there are for the horizon represented by the curve.

Figure 1b.

Comparison of ranked individual KSS – 1990: II to 2011: IV.

For all subfigures, KSS is plotted on the vertical axis, and the number of forecasters is plotted on the horizontal axis. Each point in the graph represents one forecaster. Points joined by a curve represent forecasters forecasting for the same horizon. For each horizon, forecasters are ranked based on KSS. The longer a curve is, i.e., the more points are connected by a curve, the more forecasters there are for the horizon represented by the curve.

Figure 2.

Conditional density of pooled individual forecasts.

In each subfigure, the two conditional densities of pooled individual forecasts are shown. All individuals in the SPF data are included (i.e., no participation restriction is imposed). The figure covers the entire sample period. From top to bottom are subfigures for horizons 0 to 4. The two vertical lines are the two conditional means.

Figure 3.

Benchmark simple averages and combined forecasts based on valuable individual forecasts – 1969:I to 2011:IV.

The top figure shows the benchmark simple average forecasts (i.e., the Anxious Index, dashed line) and equally weighted combined forecasts (solid line) of the real GDP decline over the next quarter ($h = 1$). The bottom figure shows the benchmark simple average forecasts (dashed line) and BLPa combined forecasts (solid line) for two-quarter-ahead forecasts ($h = 2$). The shaded areas are periods of real GDP decline.

Table 1

Descriptive statistics on individual QPS test results and skill scores.

Subsample	Horizon	Number of forecasters	QPS				Mean skill score
			Mean	Std. Dev.	Max	Min	
1969:I to 1990:I	0	51	0.124	0.043	0.049	0.225	0.383
	1	48	0.170	0.040	0.084	0.255	0.178
	2	48	0.195	0.038	0.110	0.284	0.074
	3	49	0.225	0.046	0.136	0.315	-0.037
	4	47	0.255	0.058	0.159	0.369	-0.117
1990:II to 2011:IV	0	47	0.067	0.041	0.015	0.278	0.392
	1	45	0.081	0.043	0.030	0.297	0.283
	2	44	0.101	0.049	0.041	0.263	0.120
	3	43	0.126	0.065	0.054	0.351	-0.023
	4	42	0.128	0.067	0.052	0.400	-0.074

This table gives descriptive statistics on individual forecasters' QPS and skill scores for those satisfying the participation requirement. (A naïve forecast of 14% is used to calculate skill score.)

Table 2

Results from testing the value of individual forecasts

Subsample	Horizon	Number of forecasters	Significance level			
			0.1%	1%	5%	10%
1969:I to 1990:I	0	51	13	39	50	51
	1	48	4	19	37	43
	2	48	0	4	18	27
	3	49	0	1	6	10
	4	47	0	0	4	4
1990:II To 2011:IV	0	47	1	11	26	40
	1	45	1	7	26	40
	2	44	2	9	25	31
	3	43	3	6	16	18
	4	42	2	3	11	14

This table gives the numbers of forecasters whose forecasts of a real GDP decline are considered to have value in the sense of Merton (1981), under four commonly-used significance levels. Column 2 of the table gives the total number of forecasters tested for each horizon. Each forecaster is tested independently of other forecasters.

Table 3

Estimated parameters for a beta-transformed linear combination of group forecasts.

Subsample	Horizon	Method	Log-likelihood	W_1	W_2	W_3	W_4	alpha	beta
In-sample exercise									
1969:I to 1990:I	0	OLP	-22.25	0.82	0.00	0.18	0.00	1.00	1.00
		BLP	-22.09	0.79	0.00	0.21	0.00	1.19	1.19
		BLPa	-22.08	0.78	0.00	0.22	0.00	1.18	1.14
	1	OLP	-32.24	0.00	0.00	0.56	0.44	1.00	1.00
		BLP	-31.78	0.32	0.00	0.39	0.28	1.24	1.24
		BLPa	-28.77	0.00	0.93	0.07	0.00	1.48	1.50
	2	OLP	-32.33	0.47	0.00	0.00	0.53	1.00	1.00
		BLP	-31.21	0.06	0.62	0.00	0.33	1.22	1.22
		BLPa	-31.16	0.02	0.65	0.00	0.33	1.12	1.02
1990:II To 2011:IV	0	OLP	-18.21	0.33	0.00	0.67	0.00	1.00	1.00
		BLP	-11.01	0.20	0.00	0.40	0.40	3.59	3.59
		BLPa	-8.84	0.30	0.00	0.23	0.47	3.39	1.84
	1	OLP	-22.21	0.00	0.33	0.67	0.00	1.00	1.00
		BLP	-16.66	0.03	0.00	0.97	0.00	3.08	3.08
		BLPa	-10.81	0.67	0.00	0.33	0.00	3.73	1.50
	2	OLP	-20.13	0.51	0.23	0.00	0.25	1.00	1.00
		BLP	-13.65	0.00	0.67	0.00	0.33	3.72	3.72
		BLPa	-12.43	0.00	0.67	0.00	0.33	6.64	9.87
Out-of-sample exercise									
1979:I to 1990:I	0	OLP	-12.33	0.00	0.67	0.33	0.00	1.00	1.00
		BLP	-11.76	0.40	0.40	0.20	0.00	1.00	1.00
		BLPa	-10.57	0.40	0.40	0.20	0.00	1.34	0.82
	1	OLP	-19.19	0.00	0.00	0.67	0.33	1.00	1.00
		BLP	-18.33	0.51	0.00	0.48	0.00	1.00	1.00
		BLPa	-17.36	0.00	0.33	0.00	0.67	0.67	0.29
	2	OLP	-14.62	0.23	0.32	0.45	0.00	1.00	1.00
		BLP	-14.40	0.00	0.37	0.64	0.00	1.33	1.33
		BLPa	-13.91	0.27	0.55	0.17	0.00	0.76	0.39

This table gives the log-likelihood value, estimated weights for each series of group forecast, and estimated parameters α and β for the beta-transformation. OLP: optimally weighted linear opinion pool, i.e., $\alpha = \beta = 1$; BLP: beta-transformed linear opinion pool, i.e., $\alpha = \beta = 1$; BLPa: asymmetric beta-transformed linear opinion pool, i.e., $\alpha > 0$; $\beta > 0$.

Table 4

Forecast evaluation – in-sample: 1969:I to 2011:IV.

Forecast	$E(p y=1)$	$E(p y=0)$	Odds ratio	KSS	QPS	Skill score	Calibration error	Resolution
Current-quarter forecasts ($h = 1$)								
Group1	0.663	0.161	10.247	0.502	0.077	0.360	0.017	0.061
Group2	0.639	0.163	9.094	0.476	0.084	0.302	0.017	0.054
Group3	0.629	0.133	11.102	0.497	0.073	0.396	0.011	0.059
Group4	0.623	0.113	12.994	0.510	0.069	0.424	0.015	0.066
ELP	0.638	0.142	10.645	0.496	0.069	0.423	0.011	0.062
OLP	0.645	0.136	11.512	0.509	0.065	0.456	0.012	0.067
BLP	0.554	0.059	19.937	0.495	0.058	0.519	0.002	0.065
BLPa	0.472	0.050	17.150	0.422	0.069	0.426	0.004	0.056
Overall mean	0.604	0.133	9.948	0.471	0.066	0.448	0.007	0.062
One-quarter-ahead forecasts ($h = 2$)								
Group1	0.500	0.229	3.364	0.271	0.116	0.041	0.026	0.029
Group2	0.504	0.186	4.447	0.318	0.094	0.218	0.017	0.041
Group3	0.404	0.160	3.559	0.244	0.098	0.188	0.012	0.032
Group4	0.442	0.186	3.461	0.256	0.101	0.162	0.017	0.035
ELP	0.462	0.190	3.660	0.272	0.096	0.201	0.017	0.039
OLP	0.430	0.173	3.602	0.257	0.095	0.212	0.012	0.035
BLP	0.350	0.092	5.327	0.258	0.093	0.230	0.012	0.037
BLPa	0.443	0.082	8.865	0.361	0.078	0.357	0.003	0.044
Overall mean	0.424	0.160	3.855	0.264	0.089	0.265	0.006	0.039
Two-quarter-ahead forecasts ($h = 3$)								
Group1	0.383	0.167	3.098	0.216	0.100	0.176	0.010	0.030
Group2	0.395	0.168	3.228	0.227	0.096	0.210	0.008	0.033
Group3	0.415	0.202	2.814	0.214	0.110	0.094	0.015	0.026
Group4	0.380	0.160	3.220	0.220	0.100	0.179	0.007	0.029
ELP	0.393	0.174	3.074	0.219	0.096	0.211	0.009	0.034
OLP	0.359	0.143	3.344	0.216	0.095	0.219	0.013	0.039
BLP	0.253	0.068	4.672	0.186	0.104	0.141	0.005	0.022
BLPa	0.382	0.093	5.996	0.288	0.087	0.280	0.006	0.039
Overall mean	0.290	0.167	2.033	0.123	0.105	0.130	0.004	0.021

This table gives descriptive statistics on valuable individual forecasts, grouped forecasts, combined forecasts and benchmark simple averages (overall mean). ELP: equally weighted forecasts (not transformed); OLP: optimally weighted linear opinion pool, i.e., $\alpha = \beta = 1$; BLP: beta-transformed linear opinion pool, i.e., $\alpha = \beta = 1$; BLPa: asymmetric beta-transformed linear opinion pool, i.e., $\alpha > 0$; $\beta > 0$.

Table 5

Calibration test: BLPa vs. simple average.

Horizon	Combination method	Interval	Midpoint	Frequency	Occurrence	Relative frequency	Expectation	Weight	N(0,1)	Chi square
Current-quarter forecasts	BLPa	0.00–0.049	0.025	120	2	0.02	3.00	2.93	–0.58	0.34
		0.05–0.349	0.2	27	5	0.19	5.40	4.32	–0.19	0.04
		0.35–1.000	0.675	24	17	0.71	16.20	5.27	0.35	0.12
		Total of 3		171	24			Overall	chi square	0.50
	Simple average	0.00–0.049	0.025	47	1	0.02	1.18	1.15	–0.16	0.03
		0.05–0.349	0.2	95	6	0.06	19.00	15.20	–3.33	11.12
		0.35–1.000	0.675	28	17	0.61	18.90	6.14	–0.77	0.59
		Total of 3		170	24			Overall	chi square	11.73
One-quarter-ahead forecasts	BLPa	0.00–0.149	0.075	142	9	0.06	10.65	9.85	–0.53	0.28
		0.15–0.249	0.2	10	3	0.30	2.00	1.60	0.79	0.63
		0.25–1.000	0.625	21	12	0.57	13.13	4.92	–0.51	0.26
		Total of 3		173	24			Overall	chi square	1.16
	Simple average	0.00–0.149	0.075	96	1	0.01	7.20	6.66	–2.40	5.77
		0.15–0.249	0.2	31	4	0.13	6.20	4.96	–0.99	0.98
		0.25–1.000	0.625	44	19	0.43	27.50	10.31	–2.65	7.01
		Total of 3		171	24			Overall	chi square	13.75
Two-quarter-ahead forecasts	BLPa	0.00–0.049	0.025	101	1	0.01	2.53	2.46	–0.97	0.94
		0.05–0.249	0.15	40	6	0.15	6.00	5.10	0.00	0.00
		0.25–1.000	0.625	30	16	0.53	18.75	7.03	–1.04	1.08
		Total of 3		171	23			Overall	chi square	2.02
	Simple average	0.00–0.149	0.075	82	3	0.04	6.15	5.69	–1.32	1.74
		0.15–0.349	0.25	77	16	0.21	19.25	14.44	–0.86	0.73
		0.35–1.000	0.696	14	5	0.36	9.74	2.96	–2.76	7.60
		Total of 3		173	24			Overall	chi square	10.07

This table reports the results from a calibration test of the BLPa combined forecasts and simple averages of individual forecasts for the current quarter to two quarters ahead. At each horizon, for each combination method, the table reports the intervals and their midpoints, the frequency, the occurrence of real GDP declines for periods with forecasts falling into the intervals, the relative frequency of occurrence, and the expected number of occurrences, as well as the weight of the interval and all associated test statistics. At 5%, the critical value for the overall Chi square test statistic is 5.99 for 2 degrees of freedom.

Table 6

Forecast evaluation – out-of-sample: 1979:I to 1990:I.

Forecast	$E(p y=1)$	$E(p y=0)$	Odds ratio	KSS	QPS	Skill score	Calibration error	Resolution
Current-quarter forecasts ($h = 1$)								
Group1	0.667	0.146	11.677	0.521	0.086	0.345	0.017	0.070
Group2	0.632	0.143	10.267	0.489	0.096	0.274	0.020	0.064
Group3	0.596	0.186	6.445	0.410	0.133	-0.012	0.033	0.040
Group4	0.595	0.156	7.959	0.439	0.114	0.136	0.024	0.049
ELP	0.622	0.158	8.791	0.465	0.095	0.278	0.017	0.061
OLP	0.620	0.158	8.720	0.462	0.097	0.265	0.018	0.060
BLP	0.119	0.037	3.501	0.082	0.131	0.008	0.017	0.026
BLPa	0.050	0.014	3.793	0.036	0.148	-0.123	0.026	0.017
Overall mean	0.575	0.187	5.894	0.388	0.097	0.260	0.020	0.057
One-quarter-ahead forecasts ($h = 2$)								
Group1	0.249	0.183	1.474	0.065	0.138	-0.049	0.014	0.011
Group2	0.333	0.224	1.734	0.110	0.158	-0.201	0.041	0.017
Group3	0.382	0.213	2.285	0.169	0.146	-0.112	0.039	0.027
Group4	0.298	0.227	1.451	0.072	0.167	-0.268	0.044	0.011
ELP	0.315	0.212	1.717	0.104	0.143	-0.087	0.022	0.013
OLP	0.354	0.217	1.972	0.137	0.146	-0.110	0.033	0.022
BLP	0.127	0.094	1.408	0.033	0.132	-0.001	0.006	0.009
BLPa	0.214	0.139	1.683	0.075	0.125	0.050	0.006	0.015
Overall mean	0.359	0.234	1.839	0.126	0.138	-0.050	0.020	0.016
Two-quarter-ahead forecasts ($h = 3$)								
Group1	0.351	0.215	1.969	0.136	0.164	-0.419	0.054	0.036
Group2	0.421	0.251	2.170	0.170	0.160	-0.382	0.036	0.022
Group3	0.263	0.174	1.688	0.088	0.142	-0.225	0.015	0.020
Group4	0.271	0.220	1.318	0.051	0.177	-0.532	0.038	0.007
ELP	0.326	0.215	1.768	0.111	0.148	-0.282	0.036	0.034
OLP	0.080	0.049	1.684	0.031	0.150	-0.297	0.016	0.012
BLP	0.132	0.085	1.649	0.048	0.143	-0.238	0.010	0.013
BLPa	0.072	0.047	1.565	0.025	0.151	-0.308	0.014	0.009
Overall mean	0.262	0.226	1.212	0.035	0.132	-0.140	0.023	0.009

This table gives descriptive statistics on valuable individual forecasts, grouped forecasts, combined forecasts and benchmark simple averages (overall mean). ELP: equally weighted forecasts (not transformed); OLP: optimally weighted linear opinion pool, i.e., $\alpha = \beta = 1$; BLP: beta-transformed linear opinion pool, i.e., $\alpha = \beta = 1$; BLPa: asymmetric beta-transformed linear opinion pool, i.e., $\alpha > 0$; $\beta > 0$.