Journal of Banking & Finance 72 (2016) S172–S186

Contents lists available at ScienceDirect

Journal of Banking & Finance

journal homepage: www.elsevier.com/locate/jbf

Disagreement versus uncertainty: Evidence from distribution forecasts ☆

Fabian Krüger [a,1], Ingmar Nolte [b,*]

[a] Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany
[b] Lancaster University, Bailrigg, Lancaster LA1 4YX, UK

ARTICLE INFO

Article history:
Received 12 September 2014
Accepted 26 May 2015
Available online 20 June 2015

JEL classification:
E17
C53

Keywords:
Forecasting
Survey data
Density forecasting
Disagreement
Uncertainty

ABSTRACT

We use a cross-section of economic survey forecasts to predict the distribution of US macro variables in real time. This generalizes the existing literature, which uses disagreement (i.e., the cross-sectional variance of survey forecasts) to predict uncertainty (i.e., the conditional variance of future macroeconomic quantities). Our results show that cross-sectional information can be helpful for distribution forecasting, but this information needs to be modeled in a statistically efficient way in order to avoid overfitting. A simple one-parameter model which exploits time variation in the cross-section of survey point forecasts is found to perform well in practice.

1. Introduction

Survey forecasts of macroeconomic and financial variables are available more timely and more easily than "hard" official data, and have thus become a popular source of information for forecasting (e.g. Banbura et al., 2013). Several surveys collect forecasts at the level of individual experts. The Survey of Professional Forecasters (SPF) – which we focus on in this paper – covers macroeconomic variables, as do the Bank of England's Survey of External Forecasters (SEF), the ECB's survey of professional forecasters (ECB-SPF), as well as a range of other public and commercial sources around the globe. In addition, sources like the Institutional Brokers Estimate System (I/B/E/S) cover forecasts of financial analysts about quantities like stock prices and earning per share (see e.g. Sadka and Scherbina, 2007; Nolte et al., 2014).

Individual survey forecasts often disagree with each other by substantial amounts. The present paper analyzes whether this disagreement (and, more broadly, the entire distribution of survey forecasts) can be used to predict the distribution of US macro and finance variables. As discussed below, this question generalizes an earlier literature on the interpretation of survey disagreement. We present detailed empirical evidence on our research question, and propose a simple one-parameter method in order to exploit time variation in the cross-section of survey point forecasts.

To illustrate disagreement among SPF participants, Fig. 1 summarizes 37 T-bill rate forecasts from the 2011Q1 edition of the SPF. At the time of the survey, the three-month US government bonds ("T-bill") rate stood at 0.13, but there was considerable discussion on the future stance of US monetary policy, in particular on how long the Fed would be willing to apply Quantitative Easing (The Economist, 2011). The figure displays estimates of the distribution of forecasts, separated across five target dates. As shown by the red curve, forecasters agree that the interest rate will remain on its low level in the current quarter (2011Q1; individual forecasts range from 0.05 to 0.33). For subsequent quarters, the *average* SPF forecast increases over time – this would reflect increasingly hawkish monetary policy. However, this view is quite controversial among panelists. As the target date moves farther into the future, there is more and more disagreement among panelists, finally leading to the diffuse orange distribution at the last target date (2012Q1; forecasts range from 0.15 to 2.5).

☆ This paper is a substantially revised version of a paper circulated under the title "Disagreement, Uncertainty and the True Predictive Density". We thank Michael Clements, Tilmann Gneiting, Nikolaus Hautsch, Malte Knüppel, Sandra Nolte, Frieder Mokinski, Winfried Pohlmeier, Kenneth Wallis, conference participants at the 2011 Humboldt-Copenhagen conference, ESEM 2012, and IFABS 2014, as well as seminar participants at HU Berlin and Heidelberg University for helpful comments. The first author gratefully acknowledges financial support from the Deutsche Forschungsgemeinschaft (Grant PO 375/13-1) and the European Union Seventh Framework Programme (Grant agreement no. 290976).

* Corresponding author. Tel.: +44 (0)1524 592644.
E-mail addresses: Fabian.Krueger@h-its.org (F. Krüger), I.Nolte@lancaster.ac.uk (I. Nolte).
[1] Tel.: +49 (0)6221 533285.

http://dx.doi.org/10.1016/j.jbankfin.2015.05.007
0378-4266/

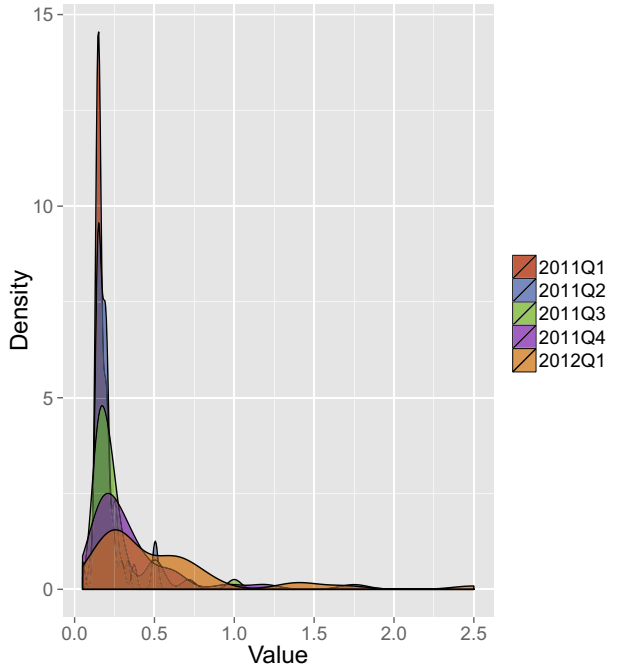

Fig. 1. T-bill rate forecasts from the 2011Q1 edition of the SPF, separated across different target dates (kernel density estimates based on 37 individual forecasts).

In macroeconomics, cross-sections of forecasts have been of interest for two main reasons. A first branch of the literature, including studies such as Lahiri and Sheng (2008), Patton and Timmermann (2010) and Coibion and Gorodnichenko (2012), aims to discern between various sources of disagreement. For example, disagreement may be generated by differential prior expectations, differential information sets, as well as different ways of interpreting a given piece of information. Distinguishing between these sources of disagreement is useful to test and improve theories of expectation formation (see Pesaran and Weale (2006) for a review). A second branch of the literature asks the practical question whether disagreement is useful to construct measures of forecast uncertainty. In an early study, Zarnowitz and Lambros (1987) note that disagreement in the cross-section of forecasters does not necessarily imply that the future is particularly uncertain. For example, each forecaster may be convinced that her point prediction would materialize with probability one. Conversely, all forecasters may agree on a particular mean forecast, but also agree that there is considerable uncertainty around this prediction. This point is lucidly summarized in the following quote:

> When the standard deviation of a set of corresponding predictions by different individuals is taken to indicate uncertainty, the underlying assumption is that this interpersonal dispersion measure is an acceptable proxy for the dispersion of intrapersonal predictive probabilities[...]. The validity of this assumption can by no means be taken for granted; it is an empirical question that is best answered by direct measurement and testing – Zarnowitz and Lambros (1987, p. 593).

Following this suggestion, many authors analyze the empirical relationship between cross sectional forecast disagreement and (various measures of) uncertainty. Using data from the Livingston Survey, Bomberger (1996) presents evidence that disagreement (measured as the cross-sectional variance of forecasts) predicts the square of subsequent forecasting errors. This suggests that disagreement may be a useful regressor in modeling conditional heteroscedasticity. Alternatively, studies such as Giordani and Söderlind (2003), Boero et al. (2008) and Lahiri and Sheng (2010) compare disagreement to variance measures constructed from histogram type survey forecasts. The latter are covered by the SPF and the SEF; they contain individual level probability forecasts for several ranges of the outcome variable. While the histograms provide rich information in principle, several conceptual and practical issues must be solved before one obtains an estimate of their implied forecast uncertainty, see Giordani and Söderlind (2003) and Engelberg et al. (2009) for careful discussions. All in all, the caveats mentioned therein make it hard to accept histogram based uncertainty measures as a "gold standard" by which the role of disagreement could be judged.

To summarize, the existing literature focuses on one cross-sectional measure (disagreement) and one notion of predictive uncertainty (variance). Here we generalize this perspective and ask whether the cross-section of point forecasts is useful to construct forecast distributions of US macroeconomic variables. That is, rather than using one variable to predict another variable, we use one distribution to predict another distribution. Our motivation is twofold. First, while variance is a natural starting point for characterizing uncertainty, the T-bill example above illustrates that other aspects like skewness or tail risk (as embodied in low quantiles) matter as well. It is thus instructive to ask whether surveys help to model forecast distributions, which by construction comprise *all* possible notions of uncertainty that may be of interest to a forecast user. Second, by using statistical performance measures for distributions, we are able to evaluate the predictive content of surveys more rigorously than studies which focus on second moments only. As detailed below, we achieve this by drawing upon a highly developed statistical literature concerned with evaluating distribution forecasts (Gneiting and Raftery, 2007). In particular, our analysis avoids the construction of a proxy for forecast variance, which has been the subject of much debate in the past.

Apart from this specific motivation, distribution forecasts have recently attracted much attention in many areas of economics and finance (e.g. Geweke and Amisano, 2011; Maheu and McCurdy, 2011). The key attraction over more traditional mean and variance forecasts is that distribution forecasts provide a full informational basis for a decision maker with arbitrary utility function. Specific examples include the distribution of product sales that is relevant to a firm, the distribution of financial returns for portfolio management, and the distribution of household income for consumption and savings decisions. In each of these examples, the econometric task is to accurately predict a probability distribution, which – together with the decision maker's utility function – prescribes an optimal decision (see Geweke, 2005, Section 2.4, for a formal discussion).

In our empirical analysis, we consider predicting the US T-bill rate, unemployment, GDP growth and inflation, using quarterly data between 1968 and 2013. Our results are based on recursive (out of sample) estimates of various forecasting models based on survey forecasts from the SPF. For GDP and inflation, which are considerably revised over time, we use macroeconomic real time data provided by the Federal Reserve Bank of Philadelphia. All of this embeds our analysis into a practically relevant forecasting scenario. This contrasts with much of the earlier literature, which considers the *in sample* relationship between disagreement and proxies of forecast variance. In order to harness information from the cross-section of SPF forecasts, we consider two distinct approaches. First, a novel "micro-level" approach uses past data and parametric assumptions to estimate the subjective forecast distribution of each individual SPF participant. We then combine all individual-level distributions to obtain a single forecast distribution. Based on just one parameter, the micro-level approach avoids overfitting, is trivial to implement and yet theoretically appealing. A second ("aggregate-level") approach rests on the opposite idea of fixing a parametric form for the predictive distribution, and feeding it with summary information from the cross-section of SPF forecasts.

The outline of the paper is as follows. Section 2 takes a more formal look at our problem, and illustrates the nexus between disagreement and uncertainty in an idealized setting. We then provide empirical evidence on our research question. To this end, Section 3 introduces the data and forecast evaluation methods, and Section 4 presents the two SPF-based models mentioned above. Section 5 considers forecast combinations in order to analyze whether surveys add information beyond state of the art time series models, rather than merely being a substitute. Section 6 provides some comparisons to histogram forecasts, and Section 7 concludes.

2. Formal motivation

This section presents a more formal motivation of disagreement under specific assumptions. Consider a group of forecasters $i = 1, \dots, n$, and suppose that the predictive density $f_{t,i}$ of forecaster i at time t has mean $\mu_{t,i}$ and variance $\sigma^2_{t,i}$.[2] Suppose further that, in order to aggregate the n forecast densities, a researcher uses the linear density combination $f_{t,a} \equiv \frac{1}{n}\sum_{i=1}^n f_{t,i}$. The assumption of equal weights $(\frac{1}{n})$ across forecasters is made purely for notational simplicity but can easily be relaxed. Well-known results for mixture distributions imply that $f_{t,a}$ has variance

$$\sigma^2_{t,a} = E_{X\sim f_{t,a}}\left[(X-\mu_{t,a})^2\right] = \underbrace{\frac{1}{n}\sum_{i=1}^n (\mu_{t,i}-\mu_{t,a})^2}_{\text{Disagreement } \mathbf{D}_t} + \underbrace{\frac{1}{n}\sum_{i=1}^n \sigma^2_{t,i}}_{\text{Average Variance } \mathbf{AV}_t}, \tag{1}$$

where $\mu_{t,a} = \frac{1}{n}\sum_{i=1}^n \mu_{t,i}$ is the average mean forecast, and $X \sim f_{t,a}$ indicates that X is distributed according to $f_{t,a}$.

The "$\mathbf{D}_t$" term in (1) is readily available from sources such as the SPF, which publishes point forecasts at the level of individual participants, for dozens of variables and at five different forecast horizons. By contrast, the "$\mathbf{AV}_t$" term in (1) is hard to estimate from surveys. While some surveys (such as SPF and SEF) cover distributional forecasts in the form of histograms, these histograms are restricted to narrow subsets of variables, time periods and forecast horizons, and require additional assumptions before yielding estimates of $\{\sigma^2_{t,i}\}$ (see Engelberg et al. (2009) as well as Section 6 below).

Given the difficulty to estimate $\mathbf{AV}_t$, it is not surprising that many authors have analyzed whether $\mathbf{D}_t$ *alone* is a useful proxy for predictive variance, see the references in the introduction. Based on (1), the role of disagreement is twofold: First, being the first summand it accounts for a certain share of predictive variance by construction. Second, $\mathbf{D}_t$ and $\mathbf{AV}_t$ may be correlated over time, in which case the former can be used as a predictor of the latter.

Interestingly, the above line of reasoning can be extended to moments beyond the second one. Standard results (e.g. Frühwirth-Schnatter, 2006, Chapter 1.2.4) imply that the third central moment of a linearly combined forecast distribution is given by

$$\begin{aligned}\lambda_{t,a} &= E_{X\sim f_{t,a}}\left[(X-\mu_{t,a})^3\right] \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^n (\mu_{t,i}-\mu_{t,a})^3}_{\text{Cross-sectional skewness } \mathbf{CSS}_t} + \overbrace{\frac{1}{n}\sum_{i=1}^n 3(\mu_{t,i}-\mu_{t,a})\sigma^2_{t,i}}^{\text{weighted Average Variance } \mathbf{wAV}_t} \\ &\quad + \underbrace{\frac{1}{n}\sum_{i=1}^n \lambda_{t,i}}_{\text{Average Skewness } \mathbf{AS}_t}, \end{aligned} \tag{2}$$

where $\lambda_{t,i} = E_{X\sim f_{t,i}}\left[(X-\mu_{t,i})^3\right]$ is the third central moment implied by the distribution of forecaster i. Note that the cross-sectional skewness (CSS) accounts for a portion of the skewness of the mixture distribution; this situation is analogous to the case of $\sigma^2_{t,a}$ above. Extending the analogy, $\mathbf{CSS}_t$ is the only term in $\lambda_{t,a}$ that can be empirically observed from survey data, and might be correlated with the other summands in (2). This motivates the idea to use $\mathbf{CSS}_t$ as a proxy for the skewness $\lambda_{t,a}$ of a forecasting distribution. This procedure could be similarly extended beyond the third moment. However, given the moderate sample size of $n \approx 40$ in the SPF, these moments of the cross sectional forecast distribution are hard to estimate in practice, and we do not consider them here.

To summarize, the cross-sectional distribution of forecasts can help to infer properties of the linear mixture $f_{t,a}$ – the latter is usually unavailable, because surveys rarely contain forecast information beyond the first moment. Of course, this motivation of disagreement hinges on the notion that the linear density combination $f_{t,a}$ is an appropriate way to aggregate individual density forecasts $\{f_{t,i}\}$ in the first place. If this assumption holds, the empirical challenge is to find a functional form which best exploits the role of disagreement. For example, consider the case of the predictive variance in Eq. (1). If $\mathbf{AV}_t$ is (approximately) a linear function of $\mathbf{D}_t$, for example, then $\sigma^2_{t,a}$ is also a linear function of $\mathbf{D}_t$.[3] This functional form assumption has been considered by Bomberger (1996).

There may be situations in which *nonlinear* combinations of individual densities are a more accurate description of reality than linear ones, so that interest lies on density combinations other than $f_{t,a}$. In these cases, the role of the cross-sectional forecast distribution is less clear. For example, Krüger (2014b) points out that in the case of *logarithmic* density combinations, the combined variance is a harmonic average of all individual variances, with disagreement not entering the equation. Even in this example, however, disagreement might be useful through its potential correlation with the (harmonic) average variance. Similar considerations apply to more complex nonlinear density combinations like the ones considered by Gneiting and Ranjan (2013). Finally, we note that histogram forecasts covered by the SPF aim to elicit probability assessments directly. In Section 6, we compare histograms to our proposed approach of reconstructing probability distributions from sets of point forecasts.

3. Data and evaluation methods

3.1. Data

As mentioned in the introduction, we consider forecasting four quarterly US macro variables: GDP growth (annualized log growth rate of real GDP), inflation (annualized log growth rate of the GDP deflator), the unemployment rate (quarterly average rate), and the T-bill rate (yield to three-month US government bonds; quarterly average rate). The GDP growth and inflation series are revised over time, which should be accounted for when designing a realistic forecasting experiment (e.g. Croushore, 2006). Whenever we estimate a model for one of the two variables, we thus employ the most recent data vintage available at that time, using real time data provided by Federal Reserve Bank of Philadelphia (2014a). For unemployment and the T-bill rate, data revisions are commonly considered too small to be of practical relevance (see e.g. Clark and Ravazzolo, 2015). For these series, we therefore use the latest data vintage as provided by Federal Reserve Bank of St. Louis (2014). Fig. 2 presents time series graphs for all series.

[2] In practice, the number of forecasters may differ across time periods, so that n becomes n_t. Furthermore, survey forecasts are specific to a certain horizon (time between forecast and realization, say, h). For simplicity, we suppress these dependencies for the moment; we use a more detailed notation when describing the models in Section 4.

[3] In formulas, if $\mathbf{AV}_t = a + b\mathbf{D}_t$ for two constants a, b, then $\sigma^2_{t,a} = a + (b+1)\mathbf{D}_t$.

In order to compute features of the cross-sectional SPF forecast distribution, we use the micro level data published by Federal Reserve Bank of Philadelphia (2014b). The data contains individual-level forecasts for GDP, inflation and unemployment from 1968:Q4 onwards, whereas coverage of the T-Bill rate starts only in 1981:Q3. The survey was initially administered by the American Statistical Association, and taken over by the Philadelphia Fed in 1990:Q2. As illustrated by Fig. 3, the number of participants varies quite substantially over time, and currently stands at around 40. The SPF survey is conducted in the middle of each quarter, which implies that the participants do not yet know the current quarter's realization. We hence refer to the current quarter forecast as horizon $h = 1$, and to the one-year ahead forecast as horizon $h = 5$ in the following.

For all models presented below, we perform parameter estimation in a rolling window fashion, with a window length of 40 quarters. The first forecast we make (=start of evaluation sample) refers to the target date 1983:Q2 for GDP, inflation, and unemployment, and to 1994:Q4 for T-Bill. These dates are chosen such that the available set of evaluation dates is the same for each forecast horizon. In order to account for possible effects of the recent crisis, we consider two evaluation samples, one ending in 2007:Q4 ("pre crisis"), and one ending in 2013:Q2 ("complete").

3.2. Evaluation methods

We consider predicting the distribution of a random variable Y_{t+h}, which denotes the stationary transform of a macro variable of interest. The forecast is based on information $\mathcal{F}_t$ available at date t, and is represented here by the probability density function $f_t(Y_{t+h})$ or, equivalently, by the cumulative distribution function $F_t(Y_{t+h})$. We next require a loss function to evaluate the quality of F_t, given a realizing outcome y_{t+h}. Throughout this paper, we interpret loss functions as penalties – the smaller, the better. A wide range of loss functions have been suggested in the literature; see e.g. Gneiting and Raftery (2007). We use the Continuous Ranked Probability Score (CRPS; Matheson and Winkler, 1976) given by

$$\mathrm{CRPS}(F_t, y_{t+h}) = \int_{\mathbb{R}} \left[F_t(z) - 1(z \geqslant y_{t+h})\right]^2 dz, \tag{3}$$

which integrates the squared distance between $F_t(z)$ and a step function $1(\cdot)$ that jumps from zero to one at $z = y_{t+h}$. It can be shown that the CRPS sets the incentive for a forecaster to reveal his true expectations about Y_{t+h} – that is, given $\mathcal{F}_t$, the *expected* score is minimized by stating the true conditional distribution.[4] Thus, the CRPS is what the literature calls a "strictly proper" loss function, and thereby qualifies as a suitable performance criterion in our context. The CRPS possesses a number of other attractive conceptual features, and has been found to be reliable in practice (cf. Gneiting and Raftery, 2007, Sections 4.2 and 8.2).

As is common in the forecasting literature, our main interest lies in *comparing* alternative forecasting methods (say, A and B) for Y_{t+h}. Following the classical framework of Diebold and Mariano (1995), this amounts to testing the null hypothesis that A and B attain the same expected CRPS (where the expectation is unconditional over time). The null hypothesis can be tested by computing the loss difference

$$d_{t+h} = \mathrm{CRPS}(F_t^A, y_{t+h}) - \mathrm{CRPS}(F_t^B, y_{t+h})$$

for a range of (quarterly) evaluation dates $t = 1, \ldots, T$, and computing the statistic

$$\frac{\bar{d}}{\widehat{V}(\bar{d})}, \tag{4}$$

where $\bar{d} = T^{-1}\sum_{t=1}^{T} d_t$ is the average CRPS difference over time, and $\widehat{V}$ denotes a (heteroscedasticity and autocorrelation) robust estimator of its variance. Below we compare forecasts in various scenarios, including comparisons of both nested and nonnested models. For simplicity, and motivated by simulation evidence in Clark and McCracken (2013), we consistently use a rectangular kernel with truncation lag $h - 1$ for the variance estimator in (4), and compare the test statistic to standard normal critical values.[5]

4. Forecasting models based on the SPF cross-section

This section first introduces two distinct forecasting approaches based on the cross-section of SPF point forecasts, and then presents empirical out-of-sample results.

4.1. Micro-level approach

Consider the cross-section of point forecasts made at date t, with target date $t+h$: $\{\mu_{t+h|t,i}\}_{i=1}^{n_t}$. The size of the SPF cross-section, n_t, may fluctuate over time as mentioned above. We propose the following simple forecast distribution:

$$f_t(Y_{t+h}) = \frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{N}(\mu_{t+h|t,i}, \sigma^2), \tag{5}$$

where $\mathcal{N}$ denotes the normal distribution and $\sigma^2 > 0$ is a single parameter to be estimated. From the discussion in Section 2, it is clear that the forecast distribution in (5) has the following properties:

- Its mean is equal to the average mean across SPF panelists, $\mu_{t+h|t,a} = \frac{1}{n_t}\sum_{i=1}^{n_t}\mu_{t+h|t,i}$.
- By Eq. (1), its variance is given by $\sigma^2_{t+h|t,a} = \underbrace{\frac{1}{n_t}\sum_{i=1}^{n_t}(\mu_{t+h|t,i} - \mu_{t+h|t,a})^2}_{\equiv \mathbf{D}_{t+h|t}} + \sigma^2$.
- By Eq. (2), its skewness is given by $\frac{1}{n_t}\sum_{i=1}^{n_t}(\mu_{t+h|t,i} - \mu_{t+h|t,a})^3 \equiv \mathbf{CSS}_{t+h|t}$.

The estimator is a simple (one-parameter) way of "scaling up" the cross-section of point forecasts to attain a justifiable forecast distribution. Specifically, the distribution in (5) inherits its first and third moments from the original cross-section of forecasts. However, and crucially, it increases the variance of the latter. The magnitude of the increase is determined empirically, by minimizing the model's training-sample CRPS with respect to $\sigma^2 > 0$. For this purpose, we use the general formula for the CRPS of a mixture of normals (Grimit et al., 2006). We also use this formula to compute the out-of-sample CRPS of the fitted model, see Appendix B for implementation details.

From an economic perspective, the model is based on the notion that all forecasters share the same degree of uncertainty (σ^2) around their prediction, and are equally informative (equal weights). While this assumption is restrictive, it simplifies the model in a number of ways. First, it makes the forecast distribution invariant to "who says what" (i.e., invariant under permutations of the ID indices $i \in \{1, \ldots, n_t\}$). This is an important practical

[4] This statement does not contradict the fact that, after y_{t+h} has realized, (3) implies that it would have been optimal to state $F_t = 1(z \geqslant y_{t+h})$.

[5] In the rare cases that the estimated variance is negative, we resort to a Newey et al. (1987) variance estimator using Bartlett weights, together with the Newey and West (1994) method for bandwidth selection, see Zeileis (2004) for implementation details in the R package "sandwich".

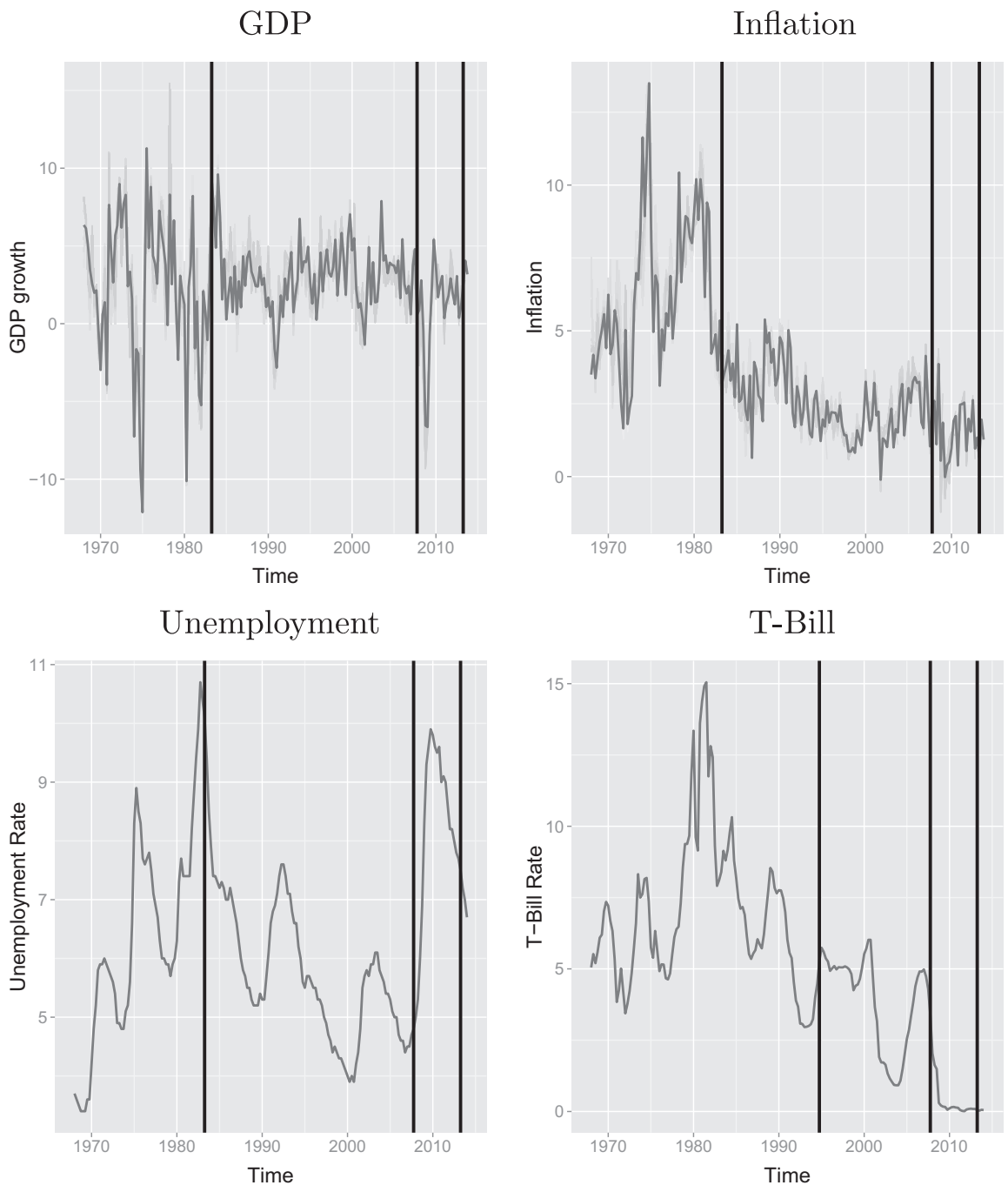

Fig. 2. Time series of the variables considered in the empirical analysis. For GDP and inflation, real time data vintages used for model estimation are plotted in light gray. In each panel, the leftmost vertical line marks the first observation of the evaluation sample. The other vertical lines mark 2007Q4 and 2013Q2, the endpoints of the two evaluation samples we consider.

advantage: The SPF data is a panel of changing composition, with frequent entry and exit of individual forecasters (Capistrán and Timmermann, 2009; Engelberg et al., 2011; D'Agostino et al., 2012). This renders estimating individual specific forecast variances or combination weights from past data very difficult. Second, the Philadelphia Fed does not guarantee the correctness of the IDs prior to 1990s, which would render individual-specific combination problematic even with complete data. Third, the assumption makes the model trivial to implement, requiring the user to estimate only a single parameter.

From a statistical perspective, the distribution in (5) corresponds to a kernel density estimate of the cross-sectional distribution of point forecasts, based on a Gaussian kernel and bandwidth parameter σ. Importantly, however, we choose σ in order to maximize the density forecasting performance of our model, which is distinct from the usual approaches toward bandwidth choice in nonparametric estimation. This difference arises because we *use* the cross-section of point forecasts to construct an accurate density forecast, rather than estimating the cross-sectional distribution of point forecasts itself. The latter would be the standard case of kernel density estimation. Furthermore, the approach in (5) can be seen as a variant of Bayesian Model Averaging used to postprocess ensembles of meteorological point forecasts (Raftery et al., 2005). However, the

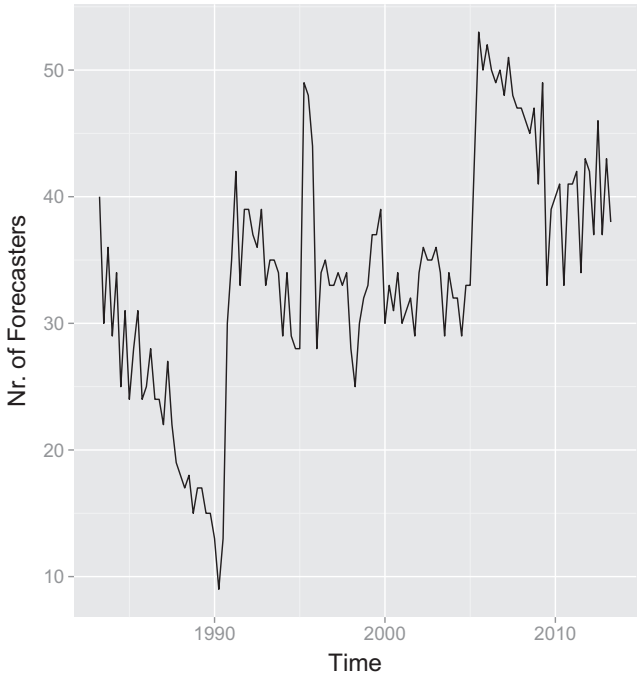

Fig. 3. Number of SPF participants (= number of forecaster IDs in the data set) over time. The numbers refer to current-quarter forecasts of inflation, but are very similar for other forecast horizons and variables.

present model is simpler in that the combination weights and variances are assumed equal across mixture components. Thus, we treat the SPF as what is called an "exchangeable ensemble" in meteorology. See Gneiting and Thorarinsdottir (2010) for further discussion on forecast ensembles in meteorology versus economics.

To illustrate the micro-level method in practice, consider current-quarter T-Bill forecasts in 2011:Q1. Here the in-sample CRPS is minimized by setting σ^2 in Eq. (5) to a very small value ($\approx 4.68 \times 10^{-5}$), thus the predictive variance $\sigma^2_{t+h|t,a}$ is dominated by the disagreement component ($\mathbf{D}_{t+h|t} \approx 2.4 \times 10^{-3}$). The top left panel of Fig. 4 presents the resulting forecast distribution for the T-Bill rate (bold line) which is an average over $n_t = 40$ Gaussians representing the individual SPF panelists (light gray). As another example, the top right panel of Fig. 4 presents the forecast distribution for the same origin date, but referring to inflation, one year ahead. Here we estimate σ^2 as approximately 1.39, which is clearly larger than disagreement ($\mathbf{D}_{t+h|t} \approx 0.78$) and thus accounts for almost two thirds of the predictive variance $\sigma^2_{t+h|t,a}$.

4.2. Aggregate-level approach

In the micro-level approach, the form of the forecast distribution is not controlled directly, but is the result of aggregating over n_t individual-level distributions. The opposite approach is to assume a parametric form for the forecast distribution, and incorporate information from the SPF cross-section via appropriate regressors. Here we consider this idea, whereby the regressors are summary statistics which represent the first three moments of the SPF cross-section for a given date, variable and forecast horizon. We use the following functional form for the h step ahead forecast density $f_t(Y_{t+h})$:

$$f_t(Y_{t+h}) = \frac{1}{\sigma_{t+h|t}} f\left(\frac{Y_{t+h} - \mu_{t+h|t}}{\sigma_{t+h|t}}; \eta, \lambda_{t+h|t}\right), \tag{6}$$

where

- $f(X; \eta, \lambda_{t+h|t})$ is the probability density function (p.d.f.) of the Hansen (1994) skewed t distribution with $\eta > 2$ degrees of freedom and skewness parameter $\lambda_{t+h|t} \in (-1, 1)$. See Appendix A for the formula of the p.d.f.
- $\mu_{t+h|t}$ and $\sigma^2_{t+h|t}$ are the conditional mean and variance of the forecast.

Importantly, note that $f(X; \eta, \lambda_{t+h|t})$ implies a mean of zero and variance of one regardless of the value for η and $\lambda_{t+h|t}$. Thus, the Hansen (1994) distribution provides a clean channel to model the forecast variance without affecting other moments of the distribution. For this reason, and because of the flexibility it provides, the distribution is our preferred choice to tackle the problem at hand.[6]

In all of the following, we fix η at a value of 20 to ease estimation; however, our results are qualitatively very robust to other plausible choice of η. As summarized in Table 1, we consider three different specifications for the remaining parameters of the distribution. Importantly, all of these forecast distributions share the same mean $\mu_{t+h|t}$ (that of the SPF forecasts) and tail thickness ($\eta = 20$). This allows us to isolate differences in forecast performance which are due to the models' predictive variance and skewness.

Specifications #1 to #3 all include survey disagreement $\mathbf{D}_{t+h|t}$ into the variance, with the additive functional form motivated by the discussion in Section 2, in particular Eq. (1). In addition to that, Specification #3 uses a measure of cross-sectional skewness to model the predictive skewness parameter $\lambda_{t+h|t}$. Our preferred measure of cross-sectional skewness is given by

$$\mathrm{sk}_{t+h|t} = \frac{\mu_{t+h|t} - \mathrm{median}_{t+h|t}}{\sqrt{\mathbf{D}_{t+h|t}}}, \tag{7}$$

where $\mathrm{median}_{t+h|t}$ is the cross-sectional median of the point forecasts. This skewness measure is standard and is similar to the one used by the Bank of England to communicate skewness in their forecast distributions for output growth and inflation (see Wallis, 2004). It appears to be more robust than using $\mathbf{CSS}_{t+h|t}$ (see Eq. 2) directly, because the latter involves estimating the third moment from a small sample of around 40 SPF point forecasts. Since $\lambda_{t+h|t}$ must be bounded between $(-1, 1)$, we model it as $\lambda_{t+h|t} = \Psi(\beta \mathrm{sk}_{t+h|t})$, where $\beta \in \mathbb{R}$ is a parameter to be estimated, and $\Psi(z) = \frac{\exp(2z)-1}{\exp(2z)+1}$ is the inverse Fisher transformation.

We estimate all specifications using maximum likelihood. For out-of-sample evaluation, we compute the CRPS of the fitted distribution by numerically calculating the integral in Eq. (3), see Appendix B for details.

The bottom row of Fig. 4 illustrates specifications #2 (solid curve) and #3 (dotted) of the aggregate approach, for the same two forecast scenarios as in the figure's top row. The two specifications yield fairly similar results for inflation (right panel), with specification #2 showing slightly more skewness. The picture is different for the T-Bill example (left panel): While specification #3 picks up the pronounced skewness of the cross-sectional forecast distribution, specification #2 – which models skewness as a fixed parameter – is very close to being symmetric.

4.3. Benchmark method

In order to evaluate the micro-level and aggregate-level approaches, we require a benchmark method. For this purpose,

[6] Focusing on US CPI inflation, Gneiting and Thorarinsdottir (2010, Section 3.1) consider an idea that is similar to our "aggregate-level" approach. However, they use a two-piece normal distribution, rather than skewed t, and focus on the first two cross-sectional survey moments.

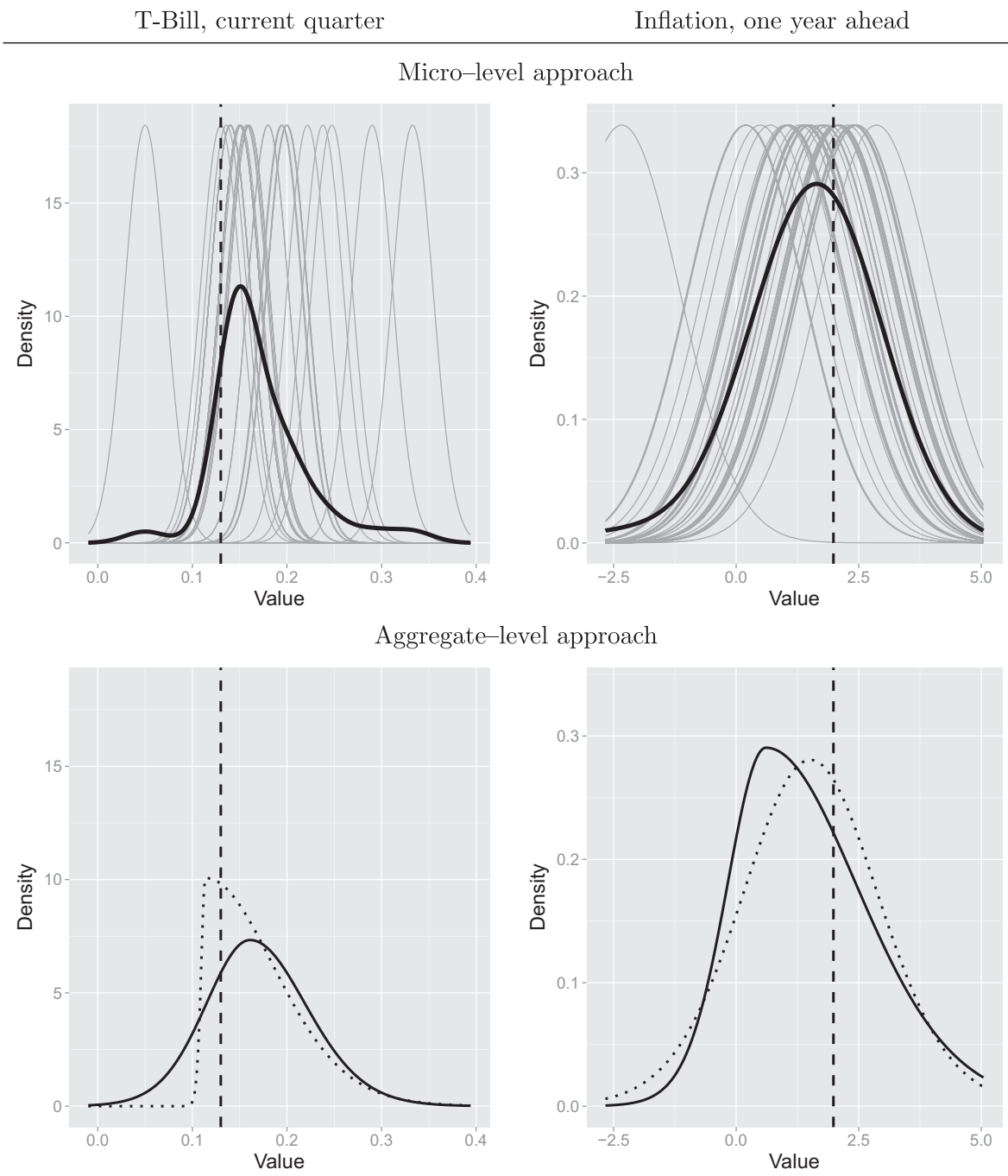

Fig. 4. Illustrative forecast distributions (origin date: 2011:Q1). Realizing value of the predictand is illustrated by a dashed vertical line. Top row: Micro-level approach, with thick line representing the mixture distribution and thin lines representing components. Bottom row: Aggregate–level approach, with solid line representing specification #2 and dotted line representing specification #3.

Table 1
Alternative parametrizations of the density in Eq. (6). All specifications set the mean equal to the SPF mean, and fix the degrees of freedom parameter at 20. α and β denote parameters to be estimated (α is forced to be strictly positive throughout). $\mathbf{D}_{t+h|t}$ and $\mathrm{sk}_{t+h|t}$ have been introduced in Section 4.1 and Eq. (7), respectively. $\Psi(\cdot)$ denotes the inverse Fisher transformation.

Specification #	$\sigma^2_{t+h\|t}$	$\lambda_{t+h\|t}$
1	$\alpha + \mathbf{D}_{t+h\|t}$	0
2	$\alpha + \mathbf{D}_{t+h\|t}$	*[constant]*
3	$\alpha + \mathbf{D}_{t+h\|t}$	$\Psi(\beta \mathrm{sk}_{t+h\|t})$

we choose the functional form in Eq. (6), with mean equal to the SPF mean, variance estimated over a rolling window of 40 observations, zero skewness, and 20 degrees of freedom as above. This benchmark makes only minimal use of the SPF cross-section, in that it incorporates the mean forecast but no other cross-sectional features like variance or skewness. It thus corresponds to a skeptical attitude towards the SPF cross-section, as expressed by some authors which emphasize conceptual differences between disagreement and uncertainty (cf. the discussion in the introduction). Note that the benchmark's rolling window based variance estimate is a simple way to approximate volatility

Table 2
Average CRPS scores, evaluation sample starting in 1983Q2 (1994Q4 for T-Bill). Specifications described in Table 1. All models estimated on rolling windows with 40 observations (=10 years of data). Dark gray coloring indicates that a method performs significantly worse than the benchmark method, whereas light gray coloring indicates significantly better performance. We use two-sided Diebold and Mariano (1995) type tests for equal CRPS and a 5% significance level; see Section 3.2 for implementation details.

Forecast horizon	Complete (⩽13Q2)					Pre crisis (⩽07Q4)				
	1	2	3	4	5	1	2	3	4	5
	GDP									
Benchmark	0.998	1.164	1.237	1.300	1.311	1.008	1.139	1.185	1.224	1.226
A-L #1	0.975	1.116	1.203	1.256	1.284	0.980	1.087	1.146	1.171	1.192
A-L #2	0.993	1.177	1.279	1.383	1.443	1.002	1.160	1.238	1.280	1.313
A-L #3	1.100	1.452	1.189	1.438	1.320	1.118	1.459	1.133	1.365	1.197
M-L	0.982	1.127	1.202	1.259	1.275	0.994	1.105	1.152	1.185	1.188
	Unemployment									
Benchmark	0.088	0.180	0.268	0.365	0.462	0.078	0.155	0.225	0.301	0.371
A-L #1	0.087	0.178	0.264	0.361	0.458	0.078	0.154	0.223	0.298	0.368
A-L #2	0.103	0.221	0.352	0.576	0.829	0.098	0.198	0.298	0.490	0.728
A-L #3	0.104	0.195	0.286	0.373	0.466	0.099	0.175	0.253	0.313	0.383
M-L	0.086	0.179	0.265	0.360	0.455	0.077	0.155	0.225	0.301	0.371
	Inflation									
Benchmark	0.485	0.541	0.591	0.636	0.681	0.476	0.542	0.594	0.645	0.693
A-L #1	0.485	0.543	0.602	0.636	0.687	0.476	0.545	0.606	0.644	0.702
A-L #2	0.561	0.577	1.111	0.866	1.053	0.568	0.585	1.225	0.924	1.144
A-L #3	0.510	0.632	0.811	0.750	1.046	0.505	0.654	0.861	0.770	1.138
M-L	0.477	0.537	0.595	0.638	0.696	0.467	0.539	0.601	0.648	0.710
	T-Bill									
Benchmark	0.080	0.249	0.435	0.629	0.820	0.080	0.239	0.402	0.576	0.737
A-L #1	0.074	0.247	0.432	0.628	0.819	0.079	0.239	0.400	0.575	0.736
A-L #2	0.112	0.247	0.434	0.629	0.875	0.132	0.240	0.439	0.644	0.835
A-L #3	0.092	0.245	0.462	0.685	0.846	0.100	0.237	0.424	0.654	0.752
M-L	0.069	0.233	0.424	0.626	0.822	0.074	0.235	0.399	0.574	0.736

clustering in macroeconomic variables. We consider more elaborate time series methods in Section 5 below.

4.4. Out-of-sample results in terms of CRPS

Table 2 presents the out-of-sample results. It suggests the following main points which are broadly consistent across both sample periods (two panels of the table).

- The micro-level ("M-L") method performs very well on the whole, and often yields significant improvements over the benchmark method at the 5% significance level; this is indicated by light gray cell coloring in Table 2. In contrast, the benchmark provides significant improvements over M-L only in one situation (inflation, $h = 5$).
- A-L #1 (disagreement-based variance, zero skewness) is the toughest overall competitor of M-L. It performs slightly better than M-L for GDP, slightly worse than M-L for T-Bill, and both methods perform very similarly for unemployment and inflation.
- A-L #1 consistently attains a smaller CRPS than the benchmark for all variables except inflation, suggesting that disagreement can be helpful in modeling the conditional variance of macro variables. Interestingly, the existing literature has mainly focused on disagreement in *inflation* forecasts.[7] Our results imply that this focus draws an overly negative picture about the usefulness of disagreement in macroeconomic forecasting.
- The A-L methods involving two parameters (#2 and 3) tend to perform poorly overall, suggesting that incorporating skewness is not helpful *in the context of the A-L methods*. This statement holds for both constant skewness (#2), and skewness modeled as a function of the SPF cross-section (#3). The poor performance of these methods may be due to the fact that the parameter estimators of these methods have to discriminate between variance and skewness of the predictive distribution. Disentangling the two may be difficult in small samples, possibly leading to estimation noise which could explain the methods' poor forecasting performance.

On the whole, we view these results as supporting the M-L approach. While A-L #1 performs similarly well empirically, it is based on the ad-hoc restriction of zero skewness which we think is conceptually undesirable. In contrast, M-L flexibly adopts to possible skewness (and other cross-sectional features). This is achieved at the small cost of fitting a single parameter, which makes the method less susceptible to overfitting than the aggregate-level methods which model skewness. To illustrate this point, the left panel of Fig. 5 plots the predictive standard deviations of the M-L and A-L #3 approaches over time, for the case of GDP and horizon $h = 4$. The standard deviations are highly correlated across the two methods. The right panel of Fig. 5 considers predictive skewness, defined as the difference between mean and median, divided by the predictive standard deviation. For M-L, skewness is very moderate. In contrast, the pronounced levels and abrupt changes of skewness in the case of A-L #3 point to possible overfitting, which may explain why the method tends to perform worse than M-L in our out-of-sample comparison.

4.5. Out-of-sample results in terms of calibration and sharpness

As pointed out by Gneiting and Katzfuss (2014), the goal of probabilistic forecasting is to produce distributions that are as sharp as possible, subject to being calibrated (i.e., in line with reality). Scoring rules like the CRPS can be seen as summary measures of both calibration and sharpness. We next illustrate how the M-L approach does a satisfactory job in terms of both aspects. To that end, consider the central 80% prediction interval, defined as the

[7] From an economic perspective, this focus is natural given the prominent role of expectations for inflation, see e.g. Mankiw et al. (2003) and Pesaran and Weale (2006).

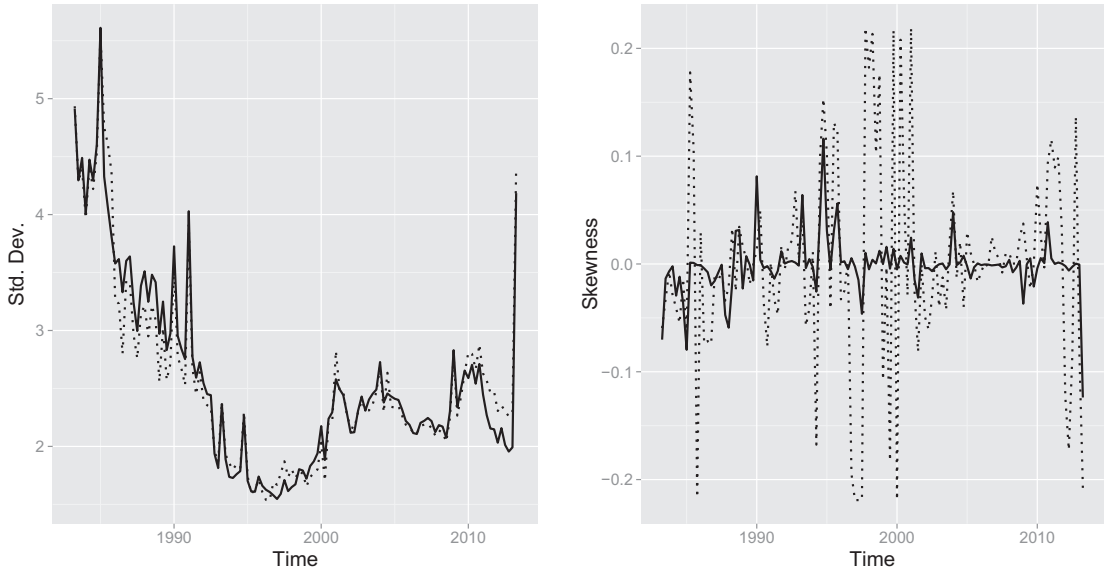

Fig. 5. Time series of predictive standard deviations (left) and skewness (right) for GDP, $h = 4$. Skewness is measured as (mean-median)/standard deviation. The solid line represents the M-L approach; the dotted line represents A-L #3.

Table 3
Coverage is defined as the share of observations that fall within the prediction intervals (the nominal target level is 80%). Length is the average length of the prediction intervals. All numbers refer to the complete sample period defined below Table 2.

Forecast horizon	1	2	3	4	5
	GDP				
Coverage (Benchmark)	0.86	0.86	0.86	0.87	0.87
Coverage (M-L)	0.83	0.87	0.86	0.84	0.83
Length (Benchmark)	5.46	6.59	6.93	7.26	7.34
Length (M-L)	5.12	6.12	6.33	6.43	6.43
	Unemployment				
Coverage (Benchmark)	0.79	0.85	0.89	0.89	0.88
Coverage (M-L)	0.74	0.83	0.83	0.87	0.88
Length (Benchmark)	0.39	0.85	1.28	1.71	2.09
Length (M-L)	0.36	0.82	1.22	1.61	1.96
	Inflation				
Coverage (Benchmark)	0.83	0.85	0.83	0.83	0.83
Coverage (M-L)	0.85	0.84	0.83	0.82	0.84
Length (Benchmark)	2.28	2.7	2.98	3.33	3.72
Length (M-L)	2.39	2.7	3.03	3.37	3.81
	T-Bill				
Coverage (Benchmark)	0.8	0.88	0.85	0.81	0.81
Coverage (M-L)	0.79	0.81	0.85	0.81	0.81
Length (Benchmark)	0.33	1.16	2.07	2.96	3.78
Length (M-L)	0.3	0.84	1.89	2.91	3.8

range between the 10% and 90% quantiles of the forecast distribution. Calibration demands that it this interval cover the actual realization with probability 0.8, whereas sharpness requires the interval to be as short as possible.

Table 3 shows that both the benchmark and M-L achieve empirical coverage rates that are fairly close to the nominal target level of 80% (ranging from 74% to 89%). The table also shows that for GDP, Unemployment and T-Bill, the M-L approach leads to shorter prediction intervals than the benchmark. For example, in the case of GDP and $h = 5$, the average length of the intervals is 7.34 for the benchmark, compared to 6.43 for M-L. Inflation is the only variable for which M-L produces slightly longer prediction intervals than the benchmark. These findings are well consistent with the results in Table 2 in terms of CRPS, in that M-L outperforms the benchmark for all variables except inflation.

5. Combination with time series forecasts

In our models based on the SPF cross-section, time-varying heteroscedasticity is generated via disagreement. We next compare these models to a flexible time series model with time varying variances, and ask whether the survey-based models contain relevant information beyond the latter. For this purpose, we consider a Bayesian Vector Autoregressive (BVAR) model with time-varying parameters and stochastic volatility, as proposed by Primiceri (2005). We then consider combining the BVAR distribution with the survey based ones, in order to analyze whether surveys can add information beyond the BVAR.

5.1. Description of the BVAR model

The BVAR postulates that the four variables of interest (say, Z_t) follow a vector autoregressive process with time-varying parameters and stochastic volatility, such that

$$Z_t = \theta_t X_t + \varepsilon_t, \tag{8}$$
$$\theta_t = \theta_{t-1} + v_t, \tag{9}$$
$$\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \Omega_t), \tag{10}$$
$$A_t \Omega_t A_t' = \Sigma_t \Sigma_t', \tag{11}$$
$$vech(A_t) = vech(A_{t-1}) + \eta_t, \tag{12}$$
$$\log diag(\Sigma_t) = \log diag(\Sigma_{t-1}) + \epsilon_t, \tag{13}$$

where Z_t is a (4×1) vector stacking the four variables at a given date, $X_t = \left[1_{[1\times 4]}, \quad Z_{t-1}', \quad Z_{t-2}'\right]'$, θ_t is a parameter vector conforming to X_t, Ω_t is the (4×4) variance–covariance matrix of ε_t, A_t is a lower triangular matrix, Σ_t is a diagonal matrix with strictly positive elements, and $\{v_t, \eta_t, \epsilon_t\}$ are mean zero, homoscedastic Gaussian disturbance vectors of appropriate dimensions.

Formally, the model is a nonlinear state space model, with unobserved states $\{\theta_t, A_t, \Sigma_t\}$. Following Primiceri (2005), we use Bayesian methods for estimation and forecasting. A first motivation for this choice is simplicity: Since the model's likelihood function is a high dimensional integral, frequentist analysis (filtering of the unobserved states, maximization of the log likelihood function) is numerically challenging and requires highly specialized methods. Bayesian estimation using Markov Chain Monte Carlo

Table 4
Average CRPS scores, evaluation sample starting in 1983Q2 (1994Q4 for T-Bill). "X/Y/Z" denotes an equally weighted linear density combination of models X, Y and Z. Dark gray coloring indicates that a method performs significantly worse than the benchmark (equally weighted combination of BVAR, M-L and A-L), whereas light gray coloring indicates significantly better performance. We use two-sided Diebold and Mariano (1995) type tests for equal CRPS and a 5% significance level; see Section 3.2 for implementation details.

Forecast horizon	Complete (⩽13Q2)					Pre crisis (⩽07Q4)				
	1	2	3	4	5	1	2	3	4	5
	GDP									
BVAR/M-L/A-L	0.995	1.128	1.197	1.263	1.281	0.985	1.077	1.121	1.166	1.176
BVAR	1.237	1.289	1.316	1.369	1.387	1.153	1.165	1.173	1.222	1.244
M-L	0.982	1.127	1.202	1.259	1.275	0.994	1.105	1.152	1.185	1.188
A-L	0.975	1.116	1.203	1.256	1.284	0.980	1.087	1.146	1.171	1.192
BVAR/M-L	1.034	1.153	1.209	1.272	1.291	1.012	1.090	1.121	1.167	1.180
BVAR/A-L	1.026	1.147	1.212	1.280	1.301	1.000	1.078	1.120	1.169	1.184
M-L/A-L	0.977	1.121	1.204	1.264	1.279	0.985	1.094	1.152	1.185	1.189
	Unemployment									
BVAR/M-L/A-L	0.089	0.181	0.271	0.367	0.462	0.078	0.154	0.224	0.297	0.362
BVAR	0.118	0.217	0.321	0.428	0.529	0.097	0.173	0.247	0.319	0.382
M-L	0.086	0.179	0.265	0.360	0.455	0.077	0.155	0.225	0.301	0.371
A-L	0.087	0.178	0.264	0.361	0.458	0.078	0.154	0.223	0.298	0.368
BVAR/M-L	0.093	0.186	0.278	0.375	0.470	0.080	0.156	0.227	0.299	0.363
BVAR/A-L	0.094	0.186	0.279	0.377	0.475	0.081	0.156	0.227	0.299	0.363
M-L/A-L	0.087	0.179	0.266	0.360	0.457	0.077	0.155	0.225	0.300	0.370
	Inflation									
BVAR/M-L/A-L	0.494	0.545	0.586	0.623	0.679	0.485	0.547	0.587	0.626	0.686
BVAR	0.577	0.596	0.605	0.643	0.701	0.568	0.601	0.602	0.638	0.696
M-L	0.477	0.537	0.595	0.638	0.696	0.467	0.539	0.601	0.648	0.710
A-L	0.485	0.543	0.602	0.636	0.687	0.476	0.545	0.606	0.644	0.702
BVAR/M-L	0.507	0.551	0.585	0.625	0.681	0.497	0.554	0.585	0.626	0.684
BVAR/A-L	0.510	0.555	0.587	0.621	0.678	0.501	0.556	0.586	0.620	0.680
M-L/A-L	0.481	0.541	0.598	0.635	0.691	0.471	0.542	0.603	0.644	0.706
	T-Bill									
BVAR/M-L/A-L	0.085	0.251	0.439	0.636	0.820	0.088	0.245	0.411	0.580	0.730
BVAR	0.174	0.344	0.526	0.711	0.879	0.170	0.327	0.495	0.648	0.780
M-L	0.069	0.233	0.424	0.626	0.822	0.074	0.235	0.399	0.574	0.736
A-L	0.074	0.247	0.432	0.628	0.819	0.079	0.239	0.400	0.575	0.736
BVAR/M-L	0.097	0.260	0.450	0.647	0.829	0.099	0.255	0.423	0.589	0.736
BVAR/A-L	0.102	0.274	0.457	0.649	0.826	0.104	0.260	0.425	0.591	0.734
M-L/A-L	0.071	0.238	0.427	0.628	0.819	0.076	0.236	0.399	0.576	0.735

(MCMC) seems more convenient (see e.g. Koop and Korobilis, 2010). A second motivation for Bayesian estimation is that informative prior distributions provide a clean channel to impose structure on the model quantities, e.g. in order to limit time variation in θ_t. This seems necessary to avoid overfitting.

We estimate the model using MCMC methods, closely following the implementation in Primiceri (2005), except for the correction by Del Negro and Primiceri (forthcoming). Our prior distributions are chosen exactly as in the original source, which entails auxiliary least squares regressions for some of the parameters. We refer the reader to the original articles for details. To estimate the model, we use the R package `bvarsv` (Krüger, 2014a), which provides an R/C++ implementation. Distribution forecasts from the model in Eqs. (8)–(13) come in the form of a simulated MCMC sample. We employ kernel density estimation to construct a forecast density, see Appendix B for details.

Importantly, our real-time approach implies that forecasts with origin date t do not incorporate the vector Z_t, but only last quarter's vector Z_{t-1}. This puts the BVAR at a disadvantage relative to the SPF participants, who have access to timely intra-quarterly information at that time (e.g. industrial production indexes as a first proxy for GDP, consumer price inflation as a proxy for GDP deflator inflation, the first monthly unemployment rate, daily T-Bill rates for the first half of the month). However, precise modeling of the SPF information set requires one to deal with issues like the timing of data releases or index construction by official agencies, and has become the focus of a specialized literature (cf. Faust and Wright, 2013 Section 2.7.3). This is well beyond the scope of the current paper, so we focus on the quarterly BVAR as a (data-wise) simpler time series model.

5.2. Forecast combinations

Below we analyze combined distribution forecasts of the form

$$\frac{1}{3}\left(f_t^{M-L}(Y_{t+h}) + f_t^{A-L}(Y_{t+h}) + f_t^{BVAR}(Y_{t+h})\right),$$

which corresponds to an equally weighted linear combination of the individual models' forecast densities (Stone, 1961). Throughout this section, "A-L" denotes the "A-L #1" specification studied in Section 4. We also consider two-model pools of the form

$$\frac{1}{2}\left(f_t^{M-L}(Y_{t+h}) + f_t^{A-L}(Y_{t+h})\right).$$

Our motivation for focusing on *equal* weights is that we aim to assess the impact of including a given model into a model. This impact appears easiest to see with equal weights.[8]

Our focus on combining *probabilistic* survey versus time series forecasts is different from earlier studies which consider various types of *point* forecast combinations involving survey and time series components (e.g. Elliott and Timmermann, 2005; Faust and Wright, 2009; Wright, 2013; Frey and Mokinski, 2014). In terms of methodology, our approach is related to studies such as Hall and Mitchell (2007), Geweke and Amisano (2011) and Krüger (2014b) which analyze combinations of distribution forecasts in economic contexts.

[8] With estimated weights (e.g. Geweke and Amisano, 2011), a model could be nominally included but actually receive very little weight.

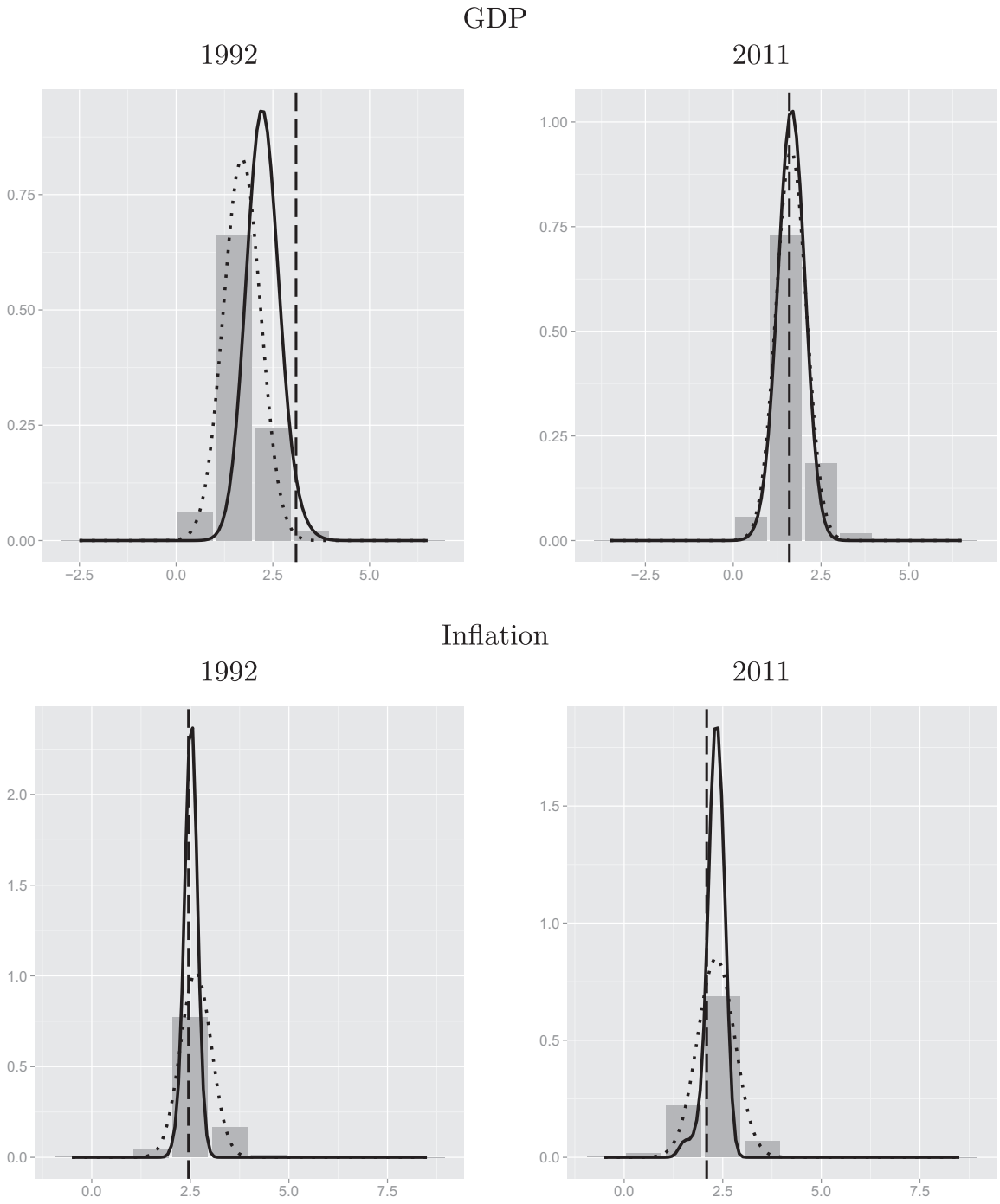

Fig. 6. Example comparisons of SPF histograms and the M-L approach. The dotted line is a normal distribution fitted to the histograms as described in the text; the solid line represents the M-L approach. The dashed vertical line marks the realizing value of the predictand.

5.3. Results

The results in Table 4 can be summarized as follows.

- For all four variables, the BVAR performs clearly worse than the SPF based methods at short horizons; this is not surprising given its smaller information set as described above. At longer horizons, the BVAR still performs somewhat worse than the survey methods, but the differences are much smaller.
- The equally weighted pool of all three models ("BVAR/M-L/A-L") clearly outperforms the BVAR. At horizons $h = 1$, the SPF based methods attain smaller (i.e., better) CRPS numbers than the pool, with differences being significant at the 5% level in some instances (light gray coloring in Table 4). At horizons $h \geqslant 2$, both M-L and A-L perform similarly to the equally weighted pool of all three methods. Hence, although the BVAR performs slightly worse than the survey methods when taken on its own, there is no penalty to including it in a

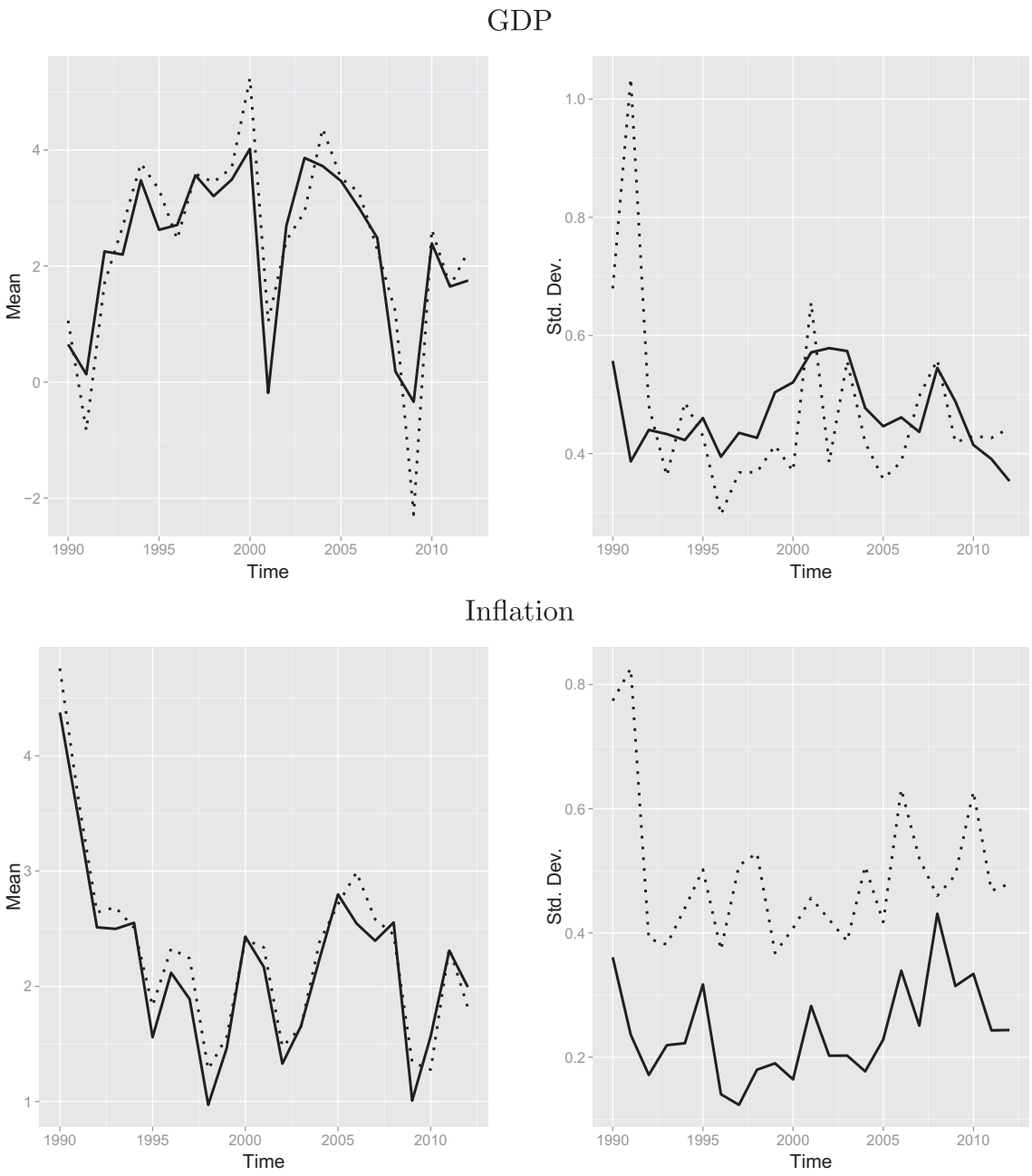

Fig. 7. Mean (left) and standard deviation (right) of the forecast distributions over time, 1990–2012. The dotted line represents a normal distribution fitted to the SPF histograms; the solid line is the M-L approach.

model pool. This finding is typical of the combination literature; Krüger (2014b) proposes an explanation in terms of the concavity of scoring rules like the CRPS.

- Among the two-model pools, the pool of the SPF based methods clearly performs best. The two pools which mix the BVAR with either M-L or A-L perform worse than the pool of all three methods, especially at short horizons.

Overall, the results in Table 4 imply that "adding" the SPF based distributions clearly improves the BVAR forecasts, whereas adding the BVAR forecast does not improve the SPF forecasting distributions. Hence the cross-section of survey forecasts is informative even when compared to (or combined with) fairly sophisticated time series forecasting models like the BVAR.

6. Comparisons to SPF histograms

We next compare the forecasting performance of the M-L approach (which stands out among the methods considered so far) to histogram forecasts provided by the SPF. Whereas M-L imposes parametric assumptions in order to construct density forecasts, the histograms aim to be direct measures of the SPF panelists' expectations. The present comparison provides further insights as to whether the simplifying assumptions made by M-L are justified from a forecasting perspective.

As mentioned in the introduction, the availability of the SPF histograms is limited to a specific "annual average" definition of the predictand, and to GDP and inflation (histogram questions for other variables have been introduced in recent years, but the

Table 5
Mean CRPS scores over time, based on observations from 1990 to 2012. The difference in mean scores is significant at the five percent level for GDP, and at the ten percent level for inflation.

	SPF histograms	M-L
GDP	0.632	0.343
Inflation	0.226	0.173

available time series are very short as of now). For the period where both point forecasts and histograms are available we carry out the following analysis. Denote by $\widetilde{Y}_a$ the *average* of the annualized log growth rates of GDP or inflation in year a. The SPF contains histogram forecasts of $\widetilde{Y}_a$ for various forecast horizons. To retain compatibility with the point forecast setup, we focus on one-step ahead histogram forecasts, made in the middle of the last quarter of year a. That is, standing in the fourth quarter (indexed by t) of year a, panelists are asked to forecast

$$\widetilde{Y}_a = \underbrace{\frac{1}{4}(Y_{t-3} + Y_{t-2} + Y_{t-1})}_{\equiv \alpha_t} + \frac{1}{4} Y_t,$$

where, as before, Y_t is the annualized log growth rate of GDP (or inflation) in quarter t. Since 1990, the timing of the SPF is such that preliminary releases of the summands in α_t are known, and the last summand Y_t needs to be predicted.[9] We can hence use the M-L approach to construct a one-step ahead forecast of Y_a, by adding α_t to the individual point forecasts and scaling the variance parameter appropriately.[10]

We follow studies such as Giordani and Söderlind (2003) and Clements and Galvão (2014) in fitting a normal distribution to the histograms, whereby the normal parameters are chosen such as to minimize the squared distance to the histogram's cumulative distribution function.[11] Fig. 6 compares the SPF histograms and the M-L approach for two illustrative dates (1992 and 2011), for both GDP and inflation. Except in one of the panels (GDP in 1992), the histogram and M-L distributions have a very similar location. Furthermore, the spread (standard deviation) of the histogram distributions exceeds that of the M-L approach for the two inflation examples (bottom panels). Fig. 7 presents a broader view on these issues, by plotting the means and standard deviations over time. The figures confirm the impression gained from the illustrations, in that the means of the histogram- versus M-L approaches are very similar, and the standard deviations of the histograms are larger than those of the M-L approach, especially for inflation. Table 5 presents CRPS scores for the histograms and the M-L approach, over the period 1990–2012 (23 observations). The M-L approach clearly outperforms the histograms for both variables.

Our reading of these results is as follows. First, the M-L approach does not necessarily provide a literal representation of the SPF panelists' probabilistic expectations. Second, this discrepancy seems desirable in the sense of leading to more accurate forecasts. These interpretations are in line with results by Clements (2014) and Clements and Galvão (2014) who document differences between subjective uncertainty measures implied by SPF histograms and ex-post measures based on realized data.

[9] As described in the documentation by Federal Reserve Bank of Philadelphia (2014c), the exact timing of the SPF prior to 1990:Q3 is less clear. We hence omit observations prior to this date from the present comparison.

[10] The variance is simply divided by 16 to reflect the fact that $\frac{1}{4}Y_t$ (rather than Y_t, as in the M-L approach) needs to be predicted.

[11] We also experimented with treating the histograms as a discrete distribution with J_t outcomes, where outcome $j \in \{1, \ldots, J_t\}$ represents the midpoint of the j-th histogram bin. This approximation yielded worse results (in terms of the histograms' forecasting performance, see below), and is hence omitted for brevity.

7. Conclusion

The present paper takes a new look at cross-sections of point forecasts, which arise in many surveys including the SPF that we focus on here. We propose to judge these cross-sections by their ability to fit distribution forecasts of macroeconomic variables, and motivate this proposal on both substantial (economic) and formal (statistical) grounds.

We find that a simple "micro-level" forecasting method based on the individual SPF point forecasts performs well in practice. This method is conceptually attractive, and can be seen as a modified version of kernel density estimation, tailored to the present goal of distribution forecasting. It is also trivial to implement, as it is based on a single parameter.

Our preferred specification assumes that survey disagreement is a useful *component* of overall predictive variance (see Section 4.1); this view is consistent with the model analyzed by Lahiri and Sheng (2010). In order to measure the usefulness of survey cross-sections, we analyze whether they help to improve distribution forecasts, as measured by statistical scoring rules. This performance metric is different from studies such as Boero et al. (2008) which judge cross-sectional survey measures by their ability to track histogram based measures of uncertainty. Our analysis in Section 6 suggests that, if the focus is on forecasting performance, matching the histogram measures may not be desirable. Of course, insofar as the histograms represent panelists' *perceived* uncertainty, they are relevant in their own right.

Appendix A. The Hansen (1994) skewed t distribution

The probability density function of the distribution is given by

$$f(X;\eta,\lambda) = \begin{cases} bc\left(1 + \frac{1}{\eta-2}\left[\frac{bX+a}{1-\lambda}\right]^2\right)^{-0.5(\eta+1)} & X < -a/b, \\ bc\left(1 + \frac{1}{\eta-2}\left[\frac{bX+a}{1+\lambda}\right]^2\right)^{-0.5(\eta+1)} & X \geqslant -a/b, \end{cases} \tag{14}$$

where

$$a = 4\lambda c\left(\frac{\eta-2}{\eta-1}\right),$$
$$b^2 = 1 + 3\lambda^2 - a^2,$$
$$c = \frac{\Gamma(0.5\times(\eta+1))}{\sqrt{\pi(\eta-2)}\Gamma(0.5\eta)}.$$

Appendix B. Computation of the CRPS

B.1. Micro-level (M-L) method

For the M-L method, the predictive distribution is a mixture of Gaussians. Grimit et al. (2006) derive an analytical expression for the CRPS in this case. Suppose the components are n Gaussians with means $\{\mu_i\}$, variances $\{\sigma_i^2\}$ and weights $\{\omega_i\}$, and an outcome y realizes. Then, the CRPS is given by

$$\mathrm{CRPS}(F,y) = \sum_{i=1}^{n}\omega_i A(y-\mu_i,\sigma_i^2) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_i\omega_j A(\mu_i-\mu_j,\sigma_i^2+\sigma_j^2), \tag{15}$$

with $A(\mu,\sigma^2) = 2\sigma\phi\left(\frac{\mu}{\sigma}\right) + \mu\left(2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right)$, where ϕ and Φ denote the p.d.f. and c.d.f. of the standard normal distribution.

B.2. Aggregate-level (A-L) method

Here the forecast distribution is skewed t, and we are not aware of a useful closed form expression for the integral in (3). We

therefore approximate the integral numerically, using the integrate function from the R software package (R Development Core Team, 2014).

B.3. BVAR

In the case of the BVAR, the forecast distribution (at a representative time point and forecast horizon) takes the form of a simulated MCMC sample of size 1000, i.e. $\mathbf{S} \equiv \{\hat{y}_i\}_{i=1}^{1000}$. We obtain this sample by running 50,000 MCMC iterations and retaining every 50th draw to reduce the autocorrelation in the sequence of draws. We then run a kernel smoother on the draws in **S**, using a Gaussian kernel and the Sheather and Jones (1991) bandwidth choice method. This yields a forecast distribution that again takes the form of a mixture of normals, thus allowing to use Eq. (15) in order to compute the CRPS.

B.4. Forecast combinations

In order to compute the CRPS of combined forecast distributions, we first approximate the A-L forecast distribution by a mixture of normals. This is done by simulating 1000 draws from the original (skewed t) distribution and running a kernel smoother, using the same procedure as for the BVAR. After this step, all three component distributions (M-L, A-L and BVAR) are mixtures of normals, and their equally weighted combination is again a mixture of normals, of the form

$$\frac{1}{3}\times\left(\frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{N}(\mu_i^{M-L},\sigma_{M-L}^2)+\frac{1}{1000}\sum_{i=1}^{1000}\mathcal{N}(\mu_i^{A-L},\sigma_{A-L}^2)+\frac{1}{1000}\sum_{i=1}^{1000}\mathcal{N}(\mu_i^{BVAR},\sigma_{BVAR}^2)\right);$$

mixtures of two of the components can be constructed in a similar way. Thus, we can again exploit the formula in (15) to compute the CRPS.

References

Banbura, M., Giannone, D., Modugno, M., Reichlin, L., 2013. Now-casting and the real-time data flow. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting, vol. 2. Elsevier, pp. 195–237.

Boero, G., Smith, J., Wallis, K.F., 2008. Uncertainty and disagreement in economic prediction: the Bank of England survey of external forecasters. Econ. J. 118, 1107–1127.

Bomberger, W.A., 1996. Disagreement as a measure of uncertainty. J. Money Credit Banking 28, 381–392.

Capistrán, C., Timmermann, A., 2009. Forecast combination with entry and exit of experts. J. Bus. Econ. Stat. 27, 428–440.

Clark, T.E., McCracken, M.W., 2013. Advances in forecast evaluation. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting, vol. 2. Elsevier, pp. 1107–1201.

Clark, T.E., Ravazzolo, F., 2015. Macroeconomic forecasting performance under alternative specifications of time-varying volatility. J. Appl. Econometrics 30, 551–575.

Clements, M.P., 2014. Forecast uncertainty–ex ante and ex post: US inflation and output growth. J. Bus. Econ. Stat. 32, 206–216.

Clements, M.P., Galvão, A.B., 2014. Measuring macroeconomic uncertainty: US inflation and output growth. Working Paper, University of Reading and University of Warwick. <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2436810> (accessed 24.06.2015).

Coibion, O., Gorodnichenko, Y., 2012. What can survey forecasts tell us about information rigidities? J. Polit. Econ. 120, 116–159.

Croushore, D., 2006. Forecasting with real-time macroeconomic data. In: Elliott, G., Granger, C., Timmermann, A. (Eds.), Handbook of Economic Forecasting, vol. 1. Elsevier, pp. 961–982.

D'Agostino, A., McQuinn, K., Whelan, K., 2012. Are some forecasters really better than others? J. Money Credit Banking 44, 715–732.

Del Negro, M., Primiceri, G.E., forthcoming. Time-varying structural vector autoregressions and monetary policy: a corrigendum. Rev. Econ. Stud.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econ. Stat. 13, 253–263.

Elliott, G., Timmermann, A., 2005. Optimal forecast combination under regime switching. Int. Econ. Rev. 46, 1081–1102.

Engelberg, J., Manski, C.F., Williams, J., 2009. Comparing the point predictions and subjective probability distributions of professional forecasters. J. Bus. Econ. Stat. 27, 30–41.

Engelberg, J., Manski, C.F., Williams, J., 2011. Assessing the temporal variation of macroeconomic forecasts by a panel of changing composition. J. Appl. Econometrics 26, 1059–1078.

Faust, J., Wright, J.H., 2009. Comparing Greenbook and reduced form forecasts using a large realtime dataset. J. Bus. Econ. Stat. 27, 468–479.

Faust, J., Wright, J.H., 2013. Forecasting inflation. In: Elliott, G., Timmermann, A. (Eds.), Handbook of Economic Forecasting, vol. 2. Elsevier, pp. 2–56.

Federal Reserve Bank of Philadelphia, 2014. Real-time data set for macroeconomists. <http://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/> (accessed 13.02.2014).

Federal Reserve Bank of Philadelphia, 2014. Survey of Professional Forecasters. <http://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/> (accessed 13.02.2014).

Federal Reserve Bank of Philadelphia, 2014. Survey of Professional Forecasters – documentation. <http://www.philadelphiafed.org/research-and-data/real-time-center/survey-of-professional-forecasters/> (accessed 18.04.2015).

Federal Reserve Bank of St. Louis, 2014. Federal reserve economic data. <http://research.stlouisfed.org/fred2/> (accessed 30.07.2014).

Frey, C., Mokinski, F., 2014. Forecasting with Bayesian vector autoregressions estimated using professional forecasts. Working Paper, University of Konstanz.

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer.

Geweke, J., 2005. Contemporary Bayesian Econometrics and Statistics. John Wiley & Sons.

Geweke, J., Amisano, G., 2011. Optimal prediction pools. J. Econometrics 164, 130–141.

Giordani, P., Söderlind, P., 2003. Inflation forecast uncertainty. Eur. Econ. Rev. 47, 1037–1059.

Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annu. Rev. Stat. Appl. 1, 125–151.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. 102, 359–378.

Gneiting, T., Ranjan, R., 2013. Combining predictive distributions. Electron. J. Stat. 7, 1747–1782.

Gneiting, T., Thorarinsdottir, T.L., 2010. Predicting inflation: professional experts versus no-change forecasts. Working Paper, Heidelberg Institute for Theoretical Studies. <http://arxiv.org/pdf/1010.2318v1.pdf> (accessed: 20.02.2014).

Grimit, E.P., Gneiting, T., Berrocal, V.J., Johnson, N.A., 2006. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Quart. J. Roy. Meteorol. Soc. 132, 2925–2942.

Hall, S.G., Mitchell, J., 2007. Combining density forecasts. Int. J. Forecasting 23, 1–13.

Hansen, B.E., 1994. Autoregressive conditional density estimation. Int. Econ. Rev. 35, 705–730.

Koop, G., Korobilis, D., 2010. Bayesian multivariate time series methods for empirical macroeconomics. Foundations Trends Econometrics 3, 267–358.

Krüger, F., 2014. bvarsv: Bayesian analysis of a vector autoregressive model with stochastic volatility and time-varying parameters, R package version 1.0.

Krüger, F., 2014. Combining density forecasts under various scoring rules: an analysis of UK inflation. Working Paper, Heidelberg Institute for Theoretical Studies. <https://sites.google.com/site/fk83research/papers> (accessed: 14.04.2015).

Lahiri, K., Sheng, X., 2008. Evolution of forecast disagreement in a Bayesian learning model. J. Econometrics 144, 325–340.

Lahiri, K., Sheng, X., 2010. Measuring forecast uncertainty by disagreement: the missing link. J. Appl. Econometrics 25, 514–538.

Maheu, J.M., McCurdy, T.H., 2011. Do high-frequency measures of volatility improve forecasts of return distributions? J. Econometrics 160, 69–76.

Mankiw, N.G., Reis, R., Wolfers, J., 2003. Disagreement about inflation expectations. NBER Macroecon. Annu. 18, 209–248.

Matheson, J.E., Winkler, R.L., 1976. Scoring rules for continuous probability distributions. Manage. Sci. 22, 1087–1096.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Newey, W.K., West, K.D., 1994. Automatic lag selection in covariance matrix estimation. Rev. Econ. Stud. 61, 631–653.

Nolte, I., Nolte, S., Vasios, M., 2014. Sell-side analysts' career concerns during banking stresses. J. Banking Finance 49, 424–441.

Patton, A.J., Timmermann, A., 2010. Why do forecasters disagree? Lessons from the term structure of cross-sectional dispersion. J. Monet. Econ. 57, 803–820.

Pesaran, M.H., Weale, M., 2006. Survey expectations. In: Elliott, G., Granger, C.W., Timmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier, pp. 715–776.

Primiceri, G.E., 2005. Time varying structural vector autoregressions and monetary policy. Rev. Econ. Stud. 72, 821–852.

R Development Core Team, 2014. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155–1174.

Sadka, R., Scherbina, A., 2007. Analyst disagreement, mispricing, and liquidity. J. Finance 62, 2367–2403.

Sheather, S., Jones, M., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Stat. Soc., Ser. B 53, 683–690.

Stone, M., 1961. The opinion pool. Ann. Math. Stat. 32, 1339–1342.
The Economist, 2011. What will the Fed do? <http://www.economist.com/blogs/freeexchange/2011/02/monetary_policy_0> (Online Commentary, February 7 2011; accessed 24.06.2015).
Wallis, K.F., 2004. An assessment of Bank of England and national institute inflation forecast uncertainties. Natl. Inst. Econ. Rev. 189, 64–71.
Wright, J.H., 2013. Evaluating real-time forecasts with an informative democratic prior. J. Appl. Econometrics 28, 762–776.
Zarnowitz, V.A., Lambros, L.A., 1987. Consensus and uncertainty in economic prediction. J. Polit. Econ. 95, 591–621.
Zeileis, A., 2004. Econometric computing with HC and HAC covariance matrix estimators. J. Stat. Software 11, 1–17.